LREC 2010 Workshop

Methods for the automatic acquisition of Language Resources and their evaluation methods

PROCEEDINGS

Edited by

Núria Bel, Béatrice Daille, Andrejs Vasiljevs

May 23, 2010

Workshop Organisers

Núria Bel
Universitat Pompeu Fabra - Barcelona, Spain

Béatrice Daille
Université de Nantes - Nantes, France

Andrejs Vasiljevs
Tilde - Riga, Latvia

Workshop Programme Committee

Victoria Arranz	ELDA - Evaluations and Language resources Distribution Agency. Paris, France
Nicoletta Calzolari	ILC-CNR, Istituto di Linguistica Computazionale Pisa, Italy
Eric de la Clergerie	INRIA – I. National de Recherche en Informatique et Automatique. France
Nancy Ide	Vassar College - Department of Computer Science New York, USA
Anna Korhonen	University of Cambridge - Natural Language and Information Processing Group (NLIP). Cambridge, United Kingdom
Montserrat Marimon	Universitat de Barcelona Barcelona, Spain
Adeline Nazarenko	LIPN - The Computer Science Lab of the Paris-Nord University. Paris, France
Stelios Piperidis	ILSP - Institute for Language Speech Processing Athens, Greece
Prokopis Prokopidis	ILSP - Institute for Language Speech Processing Athens, Greece
Valeria Quochi	ILC-CNR, Istituto di Linguistica Computazionale Pisa, Italy
Gregor Thurmair	Linguatec Munich, Germany
Anne Vilnat	LIMSI - Lab. d'Informatique Mécanique et Sciences de l'Ingénieur. Orsay Cedex, France

Preface

The FLARENET Acquisition2010 workshop wants to be the beginning of an enduring forum for the area on Automatic Acquisition and Production of Language Resources. The main objective of the workshop is the presentation of available tools and applications and especially of the evaluation methods used in order to begin to set up a common environment for the evaluation of the results of these methods and techniques.

In addition to the interest that automatic acquisition of Language Resources attracts in the academic world, its results are very close to the state of being exploited for feeding real NLP applications. In order to demonstrate the advance and usability of these acquisition methods and their results, the community has to agree on common evaluation methods, both for the assessment of the progresses achieved by each system and for the comparison of the results achieved with different methods and techniques. The workshop should also be as a discussion forum to reach a proposal to harmonize the general area with particular metadata and common vocabulary of categories to describe resources and means to acquire or produce them, as this would ease future surveys on existing tools. Finally, and even most importantly, the community has to organize the availability of common evaluation materials and the workshop wants to start this organization.

The FLARENET Working Group on Methods for the automatic construction and processing of Language Resources has launched this workshop to start the creation of such an agreement, in collaboration with the projects selected at the 4th call of the European Union 7FP that are related to this topic: ACCURAT: Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation, PANACEA: Platform for Automatic, Normalized annotation and Cost-Effective Acquisition of Language Resources for Human Language Technologies, and TTC: Terminology Extraction, Translation Tools and Comparable Corpora. The workshop will allow these projects to present their objectives and plans, as well as to discuss on the availability of their test-sets, gold-standards and other materials which can be of interest for evaluation in their projects but not only. These EU projects provide information about how to share with other researchers and LT professionals their evaluation materials to start the standardization of evaluation methods in the areas addressed.

Two keynote speakers have been invited to give their experienced opinion about evaluation from two different points of view:

Kara Warburton, with the talk “Extracting, evaluating, and preparing terminology for large-scale translation jobs” addresses the case of evaluating the results of the acquisition exercise for real-life applications. Kara Warburton is head of terminology development for IBM and is a recognized expert in terminology management. Elected chair of the LISA Terminology SIG, she is currently the project leader for the TBX submission to ISO and member of the ISO Terminology committee.

Julio Gonzalo, with the talk "Benchmarking and Evaluation Campaigns: the good, the bad and the metrics" unveils the details that lead to successful evaluation campaigns. Gonzalo is a member of the nlp.uned.es research group, where he conducts research on the application of Language Engineering to Multilingual Information Access problems and in particular in the development of evaluation metrics and methodologies. He has been involved in the coordination of CLEF (the international evaluation campaign for Multilingual Information Access applications) and WePS (Web People Search evaluation campaign).

Eight very relevant papers were selected for the poster session. The recognition and extraction of terminology is still one of the most frequent topics, and two papers, the one *Jody Foo and Magnus Merkel* and the one by Fatiha *Sadat* addressed this topic. The acquisition of lexical information is well represent first with an already classical work on sub categorization acquisition, although with an interesting application to Hungarian, by *Eszter Simon, András Serény and Anna Babarczy*; second with two papers on resources with information that is exploited in actual applications, i.e. sentiment analysis, by *Eugenie Giesbrecht* , or that has a big potential for being so, as the innovative proposal on acquiring formality lexica by *Julian Brooke, Tong Wang and Graeme Hirst*, and, third about the acquisition of metaphors by *Anna Babarczy, Ildikó Bencze M., István Fekete and Eszter Simon*. Research and evaluation on Name Entity Aligned Bilingual Corpus is represented by the work of *Xiaoyi Ma*, and work done using wikipedia resources is represented by the work of *Olivier Collin, Benoît Gaillard, Jean-Leon Bouraoui and Thomas Girault*.

The workshop also foresees a session for discussing and formulating agreements that contribute to foster the future of the automatic acquisition and production of Language Resources. A summary of the discussions will be published at the FLARENET web page www.flarenet.eu, and the conclusions and recommendation will be included in the recommendations that the network is going to propose to the relevant European decision makers for the design of new research programs and other relevant policies. Finally, we want to profit to thank the Program Committee for their collaboration and assistance in designing this interesting workshop.

Workshop's Programme

Morning session

9:30-9:45 *Welcome and Workshop Presentation*

9:45-10:30 ***Kara Warburton - "Extracting, evaluating, and preparing terminology for large-scale translation jobs"***

10.30-11.00 *Coffee Break*

10:45-11:30 ***Julio Gonzalo - "Benchmarking and Evaluation Campaigns: the good, the bad and the metrics"***

11:30-12:00 ***Andrejs Vasiļjevs - "ACCURAT****: Metrics for the evaluation of comparability of multilingual corpora"*

12:00-12:30 ***Núria Bel & Valeria Quochi - "PANACEA:*** *Evaluation in a factory of language resources and derivatives"*

12:30-13:00 ***Béatrice Daille - "TTC:*** *Evaluation procedures of multilingual terminology acquired from comparable corpora"*

13:00-14:30 ***Lunch***

Afternoon-session

14:30-16:00 ***Poster Session***

Automatic Acquisition of Hungarian Sub-categorization Frames.
By *Eszter Simon, András Serény and Anna Babarczy*

Exploiting Comparable Corpora for Building and Expanding Terminological Resources.
By *Fatiha Sadat*

Inducing Lexicons of Formality from Corpora.
By *Julian Brooke, Tong Wang and Graeme Hirst*

Semantic resource extraction from Wikipedia category lattice.
By *Olivier Collin, Benoît Gaillard, Jean-Leon Bouraoui and Thomas Girault*

The Automatic Identification of Conceptual Metaphors in Hungarian Texts: A Corpus-Based Analysis.
By *Anna Babarczy, Ildikó Bencze M., István Fekete and Eszter Simon*

Towards a Name Entity Aligned Bilingual Corpus*.*
By *Xiaoyi Ma*

Using Product Review Sites for Automatic Generation of Domain
By *Eugenie Giesbrecht*

Using machine learning to perform automatic term recognition
By *Jody Foo and Magnus Merkel*

16:00-16:30 ***Coffee Break***

16:30-17:30 ***Discussion Session*** *- A strategy for assessing the potential and future impact of acquisition techniques: criteria for the evaluation of methods*

Table of Contents

Extracting, evaluating, and preparing terminology for large translation jobs

Kara Warburton
IBM
8200 Warden Ave., Markham, Ontario, Canada L6G 1C7
kara@ca.ibm.com

Abstract

Companies that are active in global markets have demanding needs for translation services. Terminological resources enhance translation memories in order to meet those demands. Automated term extraction is the only way to build the scale of terminological resources required. Term extraction is not feasible without a process to clean the extracted terms that is at least partially automated and utilizes a range of complementary lexical resources. The entire process needs to apply concepts of recycling and avoidance of duplication in order to be cost effective and productive. The output of a term extraction process helps to build the corporate terminology database, which can subsequently be used in a range of applications.

1. Enterprise-scale translation needs

As a large global company, IBM® performs tremendous amounts of translation; according to recent estimates, the company translates about 450 million words annually, into as many as 46 languages, depending on market conditions. A given IBM product may comprise thousands of individual files in many different file formats. Translation schedules are tight, so the files are distributed to multiple translators who work in parallel but not necessarily together. Under these conditions, consistency and accuracy of key product terminology is a major challenge.

A pioneer in the use of translation memory, IBM developed and still uses its world-leading computer assisted translation (CAT) tool, TranslationManager. However, IBM recognized long ago that translation memory as a language resource could not completely address its translation needs. To continually refine the translation process, IBM has explored various language technologies such as machine translation and controlled authoring, and has developed terminology data and sophisticated workflow tools. This paper will focus on IBM's terminology management strategy, and specifically, the automatic acquisition and processing of terminology data in support of the translation process.

1. 1 Background of IBM's terminology strategy

IBM started developing terminological resources over 20 years ago. Long before there were any terminology management systems (TMS) available as commercial products, IBM researchers developed a sophisticated TMS, called TransLexis, which IBM Translation Services Centers (TSC) used to develop bilingual terminology databases. The bilingual databases are used in conjunction with translation memory to ensure consistency at the sub-segment level.

With multiple translators typically assigned to translate parts of a given product, it was understood that pre-translating the key product terms and providing those "standardized" translations to the translators would help to increase quality and minimize post-editing. This task was delegated to a staff member at each TSC, who simultaneously held the roles of project manager, translator, and terminologist.

It was soon realized, however, that the collection of source language (English) terms before translation was inefficient. Each TSC working on the same product was manually extracting terms, *if* time allowed, duplicating the work but with variable results. There was never enough time to extract and pre-translate enough terms. Furthermore, having not been involved in the development of the product, and with English as their *second* language, TSC terminologists sometimes had difficulty understanding and clarifying the meaning of terms that were ambiguous, highly technical, or poorly documented.

IBM's globalization leadership team agreed that terminology should be managed in a centralized fashion. TSC terminologists distributed around the world joined forces with two newly-appointed English terminologists, based in Toronto, to define the requirements and set up the process. Gradually, the bilngual databases were merged into one multilingual terminology database, or "termbase." After networking with terminologists from other companies, such as through the Localization Industry Standards Association, it was determined that an automated term extraction tool and handling process was key to building the type and scale of terminological resources necessary to support IBM's translation needs.

The goal was to establish an effective process for extracting and pre-translating key product terms so that a bilingual CAT dictionary of those terms could be provided to all the translators assigned to a given product *before* they started translating.

2. Term extraction and automated processing

In 2003, computational linguists at the IBM Watson Research Center completed the development of IBM's

term extraction tool, TermExt (McCord et al., 2003). In 2004, the English terminologists developed a semi-automated process for cleaning the output of TermExt, a process which was awarded a US patent.

The combined use of TermExt and the semi-automated process enables one person working only part time on the task to prepare massive amounts of high-quality English terminology resources to support the translation of IBM's products in all target languages globally. In 2009, for example, this one person sent over 160,000 terms to the TSCs in the form of nearly 3,000 CAT dictionaries -- each one customized for a specific target language -- to support the translation of over 260 products. That number -- 260 -- represents only about 40 percent of the total number of IBM products that are translated annually, so that the development and use of terminological resources to support translation in IBM has much room to grow.

The key to such impressive levels of productivity is the use of automation wherever possible, and the re-use, or recycling, of terminological and lexical resources. The system is designed to self-improve by "remembering" tasks that were done before so that they do not need to be repeated. One could call this approach "terminology memory" in support of translation memory.

2.1 Key features of TermExt

The key features of TermExt are described by McCord et al in *Terminology* (2003), so they are only summarized here. TermExt extracts terms from a single file, or a set of files as large as desired (the largest attempted so far was 1/2 million files). Essentially a rule-based system, it extracts only nouns or noun groups (sequences of words that form a noun concept and structure in the phrase). The nouns are lemmatized, that is, plural nouns are extracted in their singular form. Case differences are preserved, however, such that if a term occurs in both lower case and upper case, both are extracted, because they may correspond to different concepts (the upper case form could be a proper noun). Each extracted term is accompanied by a number, which represents the frequency by which the term was found in the scanned corpus, and a context sentence. Terms are also marked according to certain criteria: if the term is already contained in IBM's terminology database, if the term is prohibited, if the term is a possible spelling error or typing error, if the term is a possible neologism, and so forth. An exclusion list of general lexicon words is automatically applied (this idea will be explained further later). Common pre-modifiers such as "new" or "mandatory" are ignored. Strings contained within specific markup that is reserved for non-translatable text are ignored. Rules relating to case are applied to filter out even more non-translatable strings. And over 100 file formats are supported, reflecting the diverse range of file types used in the development of IBM products.

These features result in an output comprising a high proportion of translatable terms that are deemed to be "important" for the translation process, by virtue of their frequency, uniqueness, or translation difficulty. The output also identifies problems in the source text, which can then be fixed before the file is actually sent for translation. Even so, no term extraction tool is perfect, and there are always some irrelevant terms in the output, which we call "noise." To make the output as clean as possible for the TSCs, a centralized cleanup process was required.

2.2 Centralized cleanup

The TermExt output is generated by someone on the product development team, who runs TermExt against all the product files; however, the output is cleaned by the English terminologists. The cleanup process requires specialized skills and tools.

The aim of cleanup is to remove from the TermExt output everything that the translator would not find useful in the CAT dictionary. Reducing the list of terms to only those considered "important" for translators ensures that the TSC terminologists, who must translate the terms, are not given more work than necessary.

The cleanup process comprises the following steps. These steps are briefly described in this section.

1. Remove terms that already have translations in the IBM termbase for all languages translating the project.
2. Remove duplicates and near duplicates.
3. Remove general lexicon words in two passes: (a) automatically and then (b) manually.
4. Remove non-translatable strings and other "invalid" terms.
5. Review and consolidate families of terms.
6. Remove terms that already have translations in the IBM termbase for each language translating the project.
7. Convert the output to a CAT dictionary in XML format.

The languages into which a product is translated depends on market conditions and other factors, thus, each translation project involves a different set of target languages. The first stage is to remove English terms that already have translations, in the central termbase, for all the target languages in question. This step is carried out automatically using a perl routine, which compares lists of translated English terms for each target language to find what they have in common, and then removes any of those terms from the TermExt ouput.

In the next step, another perl routine identifies duplicates and so-called "near duplicates" in the TermExt file and reduces each group of such terms to one term. Duplicates can occur when two or more TermExt output files have been produced for a project, in which case the

terminologist merges them into one file before starting the cleanup. Near-duplicates are terms that differ superficially, such as in capitalization or hyphenation. A set of rules, which takes into account relative frequency, establishes which is the base form, and only that form is retained.

Words and terms that belong to the "general lexicon" are typically not needed. They include words whose translation would be obvious, and words that have little impact to the product user if they are translated inconsistently. For example, "information," or "installation process," or "requirement," are considered to be general lexicon words. These terms are removed from the TermExt file in two steps. First, a perl routine automatically removes any terms that are found in a centrally-maintained general lexicon exclusion list. Then, the terminologist goes through the file and deletes any further general lexicon words that remain. While this latter step is manual, an automatic process runs in the background; it tracks what the terminologist removes and adds those words to the general lexicon exclusion list. This ensures that a given general lexicon word is manually removed only once; henceforth it will be automatically removed by the perl routine.

The terminologist then goes through the file again, removing non-translatable strings, terms that contain spelling mistakes or typos, and anything else that she feels is "noise." TermExt removes non-translatable strings automatically when it creates the initial output, as long as they are enclosed in one of the standard markup tags that IBM uses for non-translatable strings. So most non-translatable strings never appear in the output. However, content creators don't always use standard markup, and when that happens, some non-translatable strings can "slip in" to the output. A TermExt output file that contains a lot of non-translatable strings provides a clue that the product development team in question needs further education on markup standards in IBM.

In the next step, the terminologist does a final check of the output. Usually at this point she sorts it alphabetically. This brings "families" of terms together; they are checked to determine if some of them can be eliminated. Consider the following set of terms:

configuration
dynamic memory
memory
memory configuration
memory management
memory map
memory map location
memory monitor
static memory

The following terms can be removed:

memory configuration (if we keep "configuration" and "memory")
memory management (if we keep "memory", then "memory management" is intuitive)
memory map location (if we keep "memory map")

The terms "memory map" and "memory monitor" are retained because they refer to concrete concepts. The terms "dynamic memory" and "static memory" are retained because of their binary nature; they should be translated systematically. The 10 terms have been reduced to seven. When you consider that the final TermExt output will be translated by as many as 30 TSC terminologists, sometimes even more, removing any redundancy increases productivity dramatically.

Compound terms that have the same base word are considered another type of term "family," but they are found by searching for the base form rather than alphabetically. For example, after taking into consideration the frequency of each term and the nature of the concept as expressed in the context sentence, the terminologist might decide that of the following four terms, only "impact plan" needs to be retained (she might want to ensure that "template" already figures in the output).

impact plan
template impact plan
reference impact plan
working impact plan

In the final step, a perl routine takes the cleaned TermExt output and compares it to the same list of translated terms, for each language, that was used in step one. In step one, only the terms that contained translations in the central termbase for ALL the target languages had been removed. That means that the output still contains terms that have been translated by each language, considered separately. This step removes those terms for each language. The perl routine generates a unique list of "new" (not yet translated) terms, one list for each language. If the product is translated to 30 languages, we now have 30 unique TermExt output files.

The last step converts each of the multiple TermExt output files into an XML format that can be imported into the IBM CAT tool. The files are then handed over to the TSC terminologist, who uses a "memory mining" application to search for possible translations of these new terms in the translation memory. This enables many of the terms to be translated semi-automatically from translation memory resources, which also guarantees a level of consistency between the terminology resources and the memories. The TSC terminologist translates the remaining terms. When the product files arrive for translation, the now bilingual CAT dictionary of standardized product terminology is ready to be provided to the multiple translators who are

working on the project. Using this terminology in parallel with the translation memory ensures that key product terms will be translated accurately and consistently across the thousands of files in the IBM product, regardless of who is translating any given file.

Of the seven steps listed in section 2.2, only 3(b), 4, and 5 require any manual intervention by a terminologist. The sequence of the steps is designed to minimize the human cleanup effort. The seven steps result in the reduction of a TermExt file by 60 to 80 percent. Thus, a TermExt file originally comprising 1,000 term candidates will be reduced to between 200 and 400 terms (depending on the target language), and the entire process takes the terminologist about 30 minutes. The resulting files are sent to usually between 9 and 20 target language terminologists who add the translations by using further automated processes. We have cut the effort of preparing source language terminology by a factor of the number of languages involved (for example, the effort is cut tenfold for a project involving ten target languages) while vastly improving the quality of the source language terminology resources used for translating IBM products.

2.3 Backflow and recycling

After the cleaned TermExt files are translated and then used to enhance translation memory during the translation of IBM products, the process doesn't end there. The translated "dictionaries" as they are called may be further enhanced by the translator while she is translating the product, for example, to correct a dictionary term that may have been incorrectly translated by the TSC terminologist because of an ambiguous context sentence, or to add more bilingual entries to enhance the performance of the system in prompting the translator with repeat terms. Entries that are created or corrected by the translator are approved by the TSC terminologist and then shared with the other translators involved in the project. When the project is finished, the translators return the now updated bilingual dictionaries to the TSC terminologist, who does a final review and sends a merged file to another terminologist who imports it into the central termbase. Importing terminology to an existing termbase is a complex technical task requiring specialized skills. Allocating this type of "mechanical" work to certain individuals enables the TSC terminologists to focus on tasks that demand their language skills, such as translating terminology, verifying translation memories, and post-editing.

The import of bilingual terminology resulting from the TermExt pipeline into the central termbase is referred to as "backflow," because this terminology ultimately ends up back at the starting point. At certain intervals, the importer re-exports terminology from the termbase and sends it to the English terminologists; specifically, for each target language, he exports a list of English nouns that have translations. (Remember, TermExt only extracts nouns, so the list of terms used for exclusion purposes in the cleanup process must contain only nouns.) These 40 or so lists of English nouns -- a unique one for each target language -- are used in steps 1 and 6 of the cleanup process.

The general lexicon exlusion list is another example of repurposing of lexical resources in this process. As explained in section 2.2, step 3 in the cleanup process removes general lexicon words by using a locally-maintained exclusion list. This exclusion list is automatically updated with new additions each time the terminologist removes general lexicon words manually. But the local list is also repurposed in TermExt itself. It was mentioned earlier that TermExt automatically removes general lexicon words during the initial scan. To do so, it uses a general lexicon exclusion list, which is provided by the English terminologists and is built into the TermExt executable program. The process is transparent to the end user; those general lexicon words are automatically removed. So, as the English terminologists build their local file during the daily cleanup process, this valuable asset is eventually integrated into TermExt itself, at which time a new local general lexicon exclusion list can start to be built. This recycling process ensures that the burden of managing the lexical resources, and the size of the TermExt files to be processed, are minimized at the local processing site.

3. Planned evolution

When we initially deployed the term extraction process, working exclusively on nouns made sense, because it had been determined that 90 percent of the concepts requiring special attention in the translation process were indeed nouns. But six years later, several factors began to raise the importance of verbs. First, the coverage of nouns in the termbase began to look very good; perhaps it was time to extend the process to verbs. Second, a surge in recent years in the number of acquisitions in IBM led to an increased focus on harmonizing functions and user interfaces. Third, the new paradigm of agile software development encouraged less documentation by emphasizing intuitive functions, again through the user interface. Under these conditions, terminology in the user interface needs to be impeccable. Many of the concepts expressed on user interfaces are in fact verbal concepts.

In late 2009, the IBM computational linguists were re-engaged to enhance TermExt so that it would extract verbs in addition to nouns. We took the opportunity to improve the quality of the context sentences as well. User interface files typically contain very little text and therefore very poor context sentences, so we knew that poor context sentences would become a problem for the TSC terminologists if we did not find a way to improve them.

3.1 New features

At the writing of this article, the new version of TermExt was still being deployed and therefore it would be premature to discuss its results or impact on the overall terminology management program. A detailed description of the new features should be the subject of another article. We would like to briefly describe some of its features to generate discussion and feedback.

It was decided to distinguish between so-called "simple" verbs, and those that are accompanied by a preposition or a particle. Verbs that are accompanied by a particle can be of special interest to translation because the particle can influence the meaning of the verb. For example, "turn on," "lock out", and "hang up" have different meanings than "turn," lock," and "hang" respectively. In the field of computing, there are often interesting combinations of verbs and prepositions or particles, such as "log on," and "check in." In sentences where the verb and preposition or particle are discontiguous, the parts had to be re-assembled in the extracted term. For example:

```
create from < vpp < 7 < It is the model from
which instances will be created.
```

Like nouns, verbs are lemmatized, including gerunds and passives. Modal verbs (can, may, have, etc.) are ignored.

Improving context sentences is achieved in three ways. First, increase the number of extracted context sentences, thereby increasing the amount of information available to the terminologist to establish the meaning of a term. Second, remove terms that have very poor context sentences, and put them into a separate file, which can subsequently be ignored or used as resources permit. Third, implement rules based on linguistic patterns, and use them to "score" context sentences; only sentences that meet a minimum score are extracted.

Compared to the previous version of TermExt, more of the features are configurable by the (advanced) user, for example, the score value of each of the rules, and the minimum score threshhold. This flexibility allows settings to be changed for specific needs. For instance, by increasing the score value for a given linguistic rule, you can push any segment that exhibits this rule to the top of the list of extracted context sentences. This technique can be used to extract, for example, all sentences that exhibit a specific pattern, such as "is-a," indicative of hyperonomy and hyponymy. Similarly, by setting the score for a certain rule to zero, the rule is disabled completely. Assigning high scores to specific patterns while ignoring others is a way to focus on specific types of terminological patterns, such as conceptual relations, while minimizing non-relevant information.

3.2 Impacts to existing tools and processes

Adding verbs to the TermExt output had a major impact on downstream processes, such as the cleanup process and translation. All the perl routines, exclusion lists, XML formats, memory mining tools, termbase import routines, in short, any process or tool that used a TermExt ouput file, had to be adapted to handle the additional part of speech values, the additional context sentences, and the additional "terms with poor contexts" file. The cleanup process itself had to be adapted, for example, it appears that a higher proportion of verbs, as compared to nouns, need to be removed. A number of software developers had to be engaged to make the changes, over and above the original computational linguists who modified TermExt itself. The experience showed that making what may appear to be minor changes to a natural language processing (NLP) tool can have a significant impact on downstream processes and tools.

4. Future prospects

As stated earlier, in spite of its benefits, IBM's term extraction process has yet to be used for translating over half of IBM's products. Fixed funding models and traditional workflows have made it difficult to shift resources from traditional translation to "terminology empowered" translation. For instance, TSC terminologists are also translators and project managers; translating an urgent job may take precedence over translating a TermExt file for another translation job. We still have some way to go toward recognizing target language terminology development as a key, distinct part of the translation process.

Nevertheless, the term extraction process has grown steadily since its inception in 2004, and will continue to do so as more products come on board. The process is also useful for integrating and harmonizing terminology from acquisitions.

As IBM's terminology process and tools evolve and our resources grow, we have been exploring additional uses for our terminological holdings outside of their traditional use in supporting translation, such as controlled authoring and search engine optimization. These explorations have been made possible by partnering with other NLP technologies such as the IBM LanguageWare® lexical analysis technology, the IBM Omnifind search engine, and and the controlled authoring software Acrolinx IQ.

5. Conclusion

The constant flow of structured terminology data from authoring to translation and back again has contributed to the exponential growth of IBM's terminology holdings. In the past four years, the number of translations in the IBM

termbase has grown from 350,000 to 650,000. This growth would not have been humanly possible given the size of our staff. Such growth can only be achieved through the use of technology to automate, as much as possible, the extraction, processing, and translation of terminology. IBM terminologists use a mix of NLP tools and language resources, including TermExt, language-based perl scripts, file format filters, memory mining tools, file checking and comparison routines, XSL processes, and advanced flexible importing and exporting functions. The continuous recycling of lexical resources in these tools is critical to their operation.

References

McCord, M., Bernth, A., Warburton, K. (2003). Terminology Extraction for Global Content Management. *Terminology*, 9.1.

Automatic Acquisition of Hungarian Subcategorization Frames

Eszter Simon, András Serény, Anna Babarczy

Cognitive Science Department, Budapest University of Technologies
H-1111 Budapest, Stoczek u. 2.
Research Institute for Linguistics, Hungarian Academy of Sciences
H-1068 Budapest, Benczúr u. 33.
eszter@nytud.hu, andras.sereny@googlemail.com, babarczy@cogsci.bme.hu

Abstract

Certain linguistic phenomena (e.g., ambiguity) pose computational problems that can only be solved if the system has access to lexical knowledge in general and to verb subcategorization frames in particular. Automatic subcategorization learning systems have been developed for most European languages. In this paper a Hungarian adaptation of successful models constructed for other languages is discussed with special emphasis on grammatical case based learning. The key approach adopted for our model is a statistical learning mechanism originally devised by Brent (Brent, 1993) and applied in a number of systems. In addition to and in parallel with the development of a system for the automatic acquisition of subcategorization frames, our project has the broader aim of modelling the mechanisms of child language acquisition, specifically the process of learning argument structures (subcategorization frames) from the input available to young children acquiring an agglutinative language. The outcome of our computational model was tested against child language corpora: the development curves characterizing the machine learning algorithms match the characteristic U-shaped acquisition curves observed in child language.

1. Introduction

A major part of the literature on natural language processing concerns the machine acquisition of some form of lexical knowledge. By the acquisition of lexical knowledge we mean the learning of the inventory of words and their idiosyncratic (unpredictable) properties (e.g. syntactic category, semantic properties). One of the lexical properties of a verb is its subcategorization frame, that is, the set of syntactic constituents that the verb can take as its complements. This knowledge is essential for the purposes of sentence production as well as processing. For instance, only by exploiting lexical information is it possible to resolve attachment ambiguities. Ambiguity appears at every stage in language processing and the lexicon plays an important role in its resolution. For example, the following sentence admits two interpretations which correspond to two different syntactic structures: *Salespeople sold the dog biscuits*. The source of ambiguity is that the verb *sell* has (at least) two frames: *sell + indirect object + direct object* and *sell + direct object*. In the first case salespeople sold biscuits to the dog, in the second case salespeople sold the small hard biscuits fed to dogs. However, for the sentence *Salespeople gave the dog biscuits* only the analogue of the first interpretation is possible since the verb *give* has no general frame *give + direct object*; in this case, a potential ambiguity is resolved in view of our knowledge of frames.

Corpus-based argument structure retrieval is a major research topic mostly applied to English (e.g. Schulte im Walde (2008)) but also to several other European languages (Ienco et al., 2008; Maragoudakis et al., 2000; Zeman and Sarkar, 2000). In our paper a Hungarian adaptation of successful models constructed for other languages is discussed. The key approach adopted for our model is a classic statistical learning mechanism originally devised by Brent (1993) and later applied in a number of systems. Our model can be seen as complementing a recent Hungarian language system (Sass, 2006), which is aimed at retrieving idiomatic, non-compositional verbal constructions containing specific lemmas. A similar problem has been addressed in (Gábor and Héja, 2007), but their main objective was the identification of semantically related verb classes, where the semantic features bore syntactic relevance in Hungarian. The models discussed in the current paper attempt to extract subcategorization frames purely on the basis of morphosyntactic features without making any reference to semantics.

The research topic of the mechanisms of lexical knowledge acquisition is also interesting in other respects. First, this is one of the key issues in psycholinguistics, and second, certain linguistic phenomena pose problems in natural language processing (NLP) that can only be resolved if the system has access to lexical knowledge. Our study both attempts to contribute to NLP research by comparing the results of child language analyses to the output of computational models, and at the same time intends to use this comparison to shed light on human learning mechanisms, specifically relevant to agglutinative languages.

Subcategorization frames, or argument frames, are defined here as the linguistic information signalling the case roles of verbal arguments. The task is therefore to decide for each verb given in the initial lexicon whether it can be mapped onto a given subcategorization frame or, more precisely, to assign a probability of a verb occurring with a given subcategorization frame. In the case of Hungarian, the primary linguistic cues defining a given subcategorization frame are morphological case markers suffixed to argument nouns.

In addition to Brent's method we have implemented two more procedures: a likelihood ratio test and a decision technique based on relative frequencies. These techniques are presented in Section 2. The methods were tested on two Hungarian corpora: in Section 3 our evaluation method and in Section 4 the results are described. Last but not least we discuss the psycholinguistic aspects of automatic subcategorization frame acquisition and the conclusions.

2. Experiments

2.1. Binomial hypothesis test

Brent was the first to use the following algorithm to extract verb frames from text corpora. Suppose we have a fixed set F of frames and a set V of verbs and for each pair $(f, v) \in F \times V$ we want to make a decision based on statistical evidence whether the verb v takes the frame f. First, for each frame $f \in F$ let us define a pattern of words and syntactic categories which indicate the presence of the frame with a high certainty. We call such form patterns *cues* for frame f. For example, the obvious cue for the English transitive verb frame might be written as VERB NP meaning that the verb must be followed by an NP in the sentence. (We shall shortly see examples of cues for Hungarian frames.) Clearly, cues are not infallible indicators of frames, hence we assign a probability of error to each cue: this is the probability that the cue appears in a sentence even though the frame does not appear in the sentence. The method requires that cues belonging to the same frame should have the same probability of error. The error probability is a variable parameter of model; the best value is determined on an empirical basis.

Once the cues have been chosen, we perform hypothesis testing to decide whether a frame f is appropriate for a verb v. Our null hypothesis is that the frame is not appropriate for the verb; we reject this null hypothesis if there is sufficient statistical evidence against it. Suppose that the verb v occurs n times in the corpus and there are $C(v, f)$ occurrences together with a cue for frame f. Now,

$$p_e = P(C(v,f) \geq m \mid v \text{ does not take } f) = \sum_{r=m}^{n} \binom{n}{r} \varepsilon_f^r (1-\varepsilon_f)^{n-r}$$

is the probability that cues for f occur m or more times together with v, where ε is the error probability for f. If p_e is smaller than a given threshold (the significance level) then we reject our null hypothesis and decide that the verb can take the frame.

Due to the variable word order characteristic of the Hungarian language, we cannot rely on exploiting particular linear configurations alone when creating the cues. The linguistic feature that our model can exploit is that Hungarian is an agglutinative language with rich morphological case marking. Besides morphological case markers suffixed to nouns, argument roles may also be signalled by postpositions. The cues used in our model are, therefore, regular expressions over the alphabet of the „KR-code", which is the coding system used for morphological annotation in the Hungarian text corpora used as the input to the learning system (Kornai et al., 2004). The regular expressions are matched against strings of morphological description. For example, the cue for the Hungarian *ditransitive* verb frame (verbs taking a complement in the accusative and a complement in the dative in either order, e.g. *ad vkinek vmit* (*give someone*-dat *something*-acc)) has the following pattern:

(CAS<ACC>.* CAS<DAT>) |
(CAS<DAT>.* CAS<ACC>)

2.2. Likelihood ratio test

Child language data suggests that following a period of correct usage, children between the ages of about 3 and 8 years tend to assume that verbs with similar meanings share an argument frame. Specifically, they generalize alternation patterns to verbs that do not allow frame alternation in the adult grammar. These errors appear to give a U-shaped learning curve, where correct usage precedes overgeneralization (see e.g., Clark (1987), Bowerman (1989), Babarczy (2002)). The explanation for the phenomenon appears to be that children first learn a few frequent expressions as unanalysed linguistic units but later extract patterns, which may then lead to overgeneralisation errors. As the conservative learning algorithm of the binomial hypothesis method does not seem to match the psycholinguistic evidence, we implemented a second method.

The likelihood ratio test is a widely used, general parametric statistical test. We apply it in the following way. Let us take a frame f and a verb v. Let I_f denote the following random variable: $I_f = 1$, if a cue for f occurs in a sentence and $I_f = 0$ otherwise; similarly, $I_v = 1$, if v appears in a sentence and $I_v = 0$ otherwise. Essentially, we would like to determine whether the random variables I_f and I_v are independent; if so, then we infer v does not take f, if not, then we infer v takes f. It is easily seen that I_f and I_v are independent if and only if the conditional distributions $I_f \mid I_v$ and $I_f \mid (1 - I_v)$ coincide. We shall use the likelihood test to make the decision. Let k_1, n_1, k_2, n_2 denote, respectively, the number of occurrences of v and a cue for f together, the number of verbs in the corpus, the number of occurrences of a cue for f with any other verb than v and the number of verb occurrences other than v. Then the logarithm of the likelihood ratio is calculated as

$$\lambda = l(\frac{k_1+k_2}{n_1+n_2}, k_1, n_1) + l(\frac{k_1+k_2}{n_1+n_2}, k_2, n_2) - \\ - l(\frac{k_1}{n_1}, k_1, n_1) - l(\frac{k_2}{n_2}, k_2, n_2),$$

where $l(q, n, k) = k \log q + (n - k) \log(1 - q)$. As is well known, λ tends toward the chi-squared distribution, so we can compare λ to critical values of χ^2 given a significance level.

Because of the sensibility of this model to the probability of the co-occurrence of a given frame with other verbs, it seems to be closer to the hypothesized human acquisition mechanisms of generalization and the subsequent gradual recovery from overgeneralized patterns.

2.3. Relative frequencies

This straightforward method was suggested by (Korhonen et al., 2000) as a baseline and it does away with the notion of significance completely. For each verb, the occurrences of cues are counted and the frames whose relative frequency of co-occurrence with the verb exceeds a threshold are selected; this threshold is determined empirically, i.e., the threshold producing the best results is selected.

3. Evaluation

3.1. Corpora

Our methods were tested on two Hungarian text corpora: the Szeged Corpus and the Hungarian Webcorpus. The Szeged Corpus (Csendes et al., 2004) is a Hungarian treebank, containing approximately 82 thousand sentences along with full morphological and syntactic annotation. For the purposes of our experiments we only retained the information concerning morphology and postpositions.

The other corpus we used is the Hungarian Webcorpus (Halácsy et al., 2004; Kornai et al., 2006), which, with over 1.48 billion words unfiltered (589 million words fully filtered), is by far the largest Hungarian language corpus. Only a section of the corpus was used here containing 832 thousand sentences. As this corpus is not annotated, we needed a part-of-speech tagger to extract the morphological information. We used the `hunpos` (Halácsy et al., 2007), which is a Hidden Markov Model-based open source part-of-speech tagger, with the Hungarian language resources of `morphdb.hu` (Trón et al., 2006).

3.2. Methodology

To measure the accuracy of a machine learning algorithm, its output has to be compared to a gold standard test dataset. The standard method of quantifying the similarity between the gold standard and the output is the CoNLL F-measure (C. J. van Rijsbergen, 1979). In the present study the gold standard is a verb list: the 1000 most frequent verbs from the Szeged Corpus and their subcategorization frames as specified by a trained linguist. According to the gold standard list, these 1000 verbs can take a total of 11 different subcategorisation frames. The evaluation method used the following procedure: a subcategorization frame was taken to be correctly assigned to a verb if the given frame was specified for this verb in the gold standard list. Based on this, precision and recall values can be calculated for the experiments. The performance of learning algorithm-based natural language processing modules is traditionally measured in precision, which is the ratio of the correct answers to the produced answers, and recall, which is the ratio of the correct answers to the total expected answers. The F-measure is, as usual, the harmonic mean of these two values.

4. Results

Our experiments used a set of different parameters: in addition to the learning algorithm, the error probability of the binomial hypothesis test and the input corpus, we also varied the number of subcategorisation frames and the number of verbs to be acquired by the model (see Table 1 below). Starting with the choice of input corpus, if we compare the results of our measurements on the two corpora we can see that even though the Webcorpus is noisy and automatic morphological parsing is a source of further errors, its sheer size outweighs these disadvantages: we obtain better results here than on the Szeged Corpus.

Looking at the different learning algorithms, we find that the likelihood ratio test gave slightly poorer results than the Brent method but the learning curve suggests that performance could be improved by using more training data (i.e., a larger corpus). Surprisingly, the best result was achieved with the method where the decision was made on the basis of relative frequency, similarly to the findings presented in Korhonen et al. (2000).

The Brent method was tested using a number of different values of error probability (shown in Column 2). The results reveal that precision improves, but recall declines with an increase in the value of error probability. The F-measure, of course, balances these values but it remains the case that lower values of error probability lead to better performance. An error probability of 0.1 gave the best results. Performance could not be improved by estimating error probabilities for individual cues.

Brent took a very cautious approach to the extraction of subcategorization frames from untagged corpora: he tried to extract just five frames. Manning (1993) extended the method by using morphological information, and also increased the number of subcategorization frames to 19. First we tested the model on all of the 11 subcategorization frames occurring in the gold standard list. If we work with the same method and parameters but reduce the learning domain to the 3 most frequent frames (*transitive, dative, ditransitive*), the F-measure considerably increases. It is a typical consequence of Zipf's Law of word distribution: a few words occur very frequently, more words occur somewhat frequently, and many words occur infrequently. This is the well-known problem of data sparseness: there is no corpus large enough to find all words at least once in it. The situation is similar in the case of the number of verbs the model is tested on. If we take into account only the 200 or 100 most frequent verbs in the evaluation, the performance of the system increases.

One of our goals was to model these learning mechanisms and compare the behaviour of the system to real-world data. To illustrate the psycholinguistic parallel we present our results graphically as well.

The model curve used here is a typical U-shaped curve observed in the acquisition of the subcategorization frames of a Hungarian verb by 3 Hungarian children (see Figure 1). The horizontal axis of the child data curve represents time (presented by Mean Length of Utterance, MLU): as input sentences accumulate, the initial conservative correct usage of constructions is overgeneralized before further input allows errors to be corrected. (The data were taken from Babarczy (2002) and restructured for the purposes of this study.)

Our results are presented graphically in Figure 2. The curve of the likelihood ratio trial shows a U-shaped curve similar to that observed in child language. The horizontal axis of the machine learning curve shows the size of the corpus, which fulfils a similar function in machine learning as the size of the input data set increases with time in child language acquisition. Although the curve rises slowly here, the U-shape is clearly visible. The recall curve of the likelihood ratio test is monotonic increasing (see Figure 3).

5. Conclusion

One of our aims in constructing a statistical learning model was to model the mechanisms of verb argument frame acquisition by young children learning a language where ar-

corpus	method	frames	verbs	precision	recall	F-measure
Szeged	bht	3	1000	63%	50%	56%
Wc	bht	3	1000	70%	67%	68%
Wc	bht 0.5	11	100	94%	34%	51%
Wc	bht 0.2	11	100	60%	71%	64%
Wc	bht 0.1	3	200	64%	94%	76%
Wc	lht	3	1000	25%	79%	39%
Wc	rel freq	3	1000	90%	67%	76%

Table 1: Results.

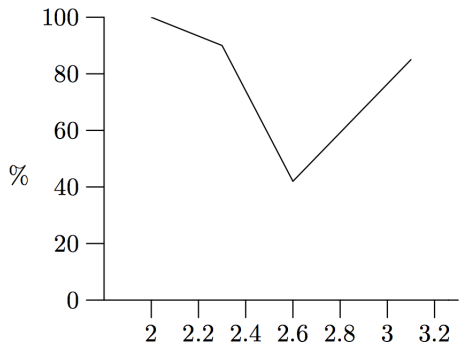

Figure 1: The ratio of correct usage of the verb *kér* by 3 Hungarian children.

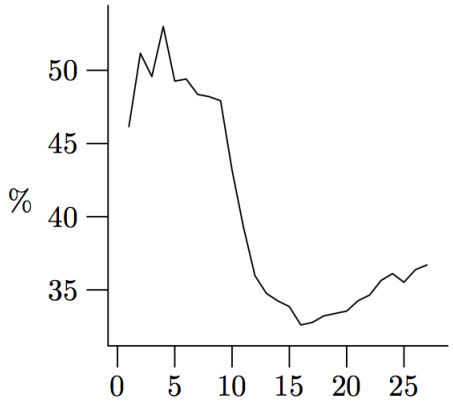

Figure 2: The precision values of likelihood ratio test.

gument roles are primarily marked by morphological cases. We have shown that data frequency and the size of the input corpus are important factors in both psycholinguistics and machine learning. Our results reveal that the performance of the system is best when a small number of highly frequent subcategorization frames need to be learnt. This pattern appears to accord with the hypothesis that the U-shape pattern characterising child language is explained by the observation that children first acquire a few very frequent constructions. At present, however, computational corpus analysis cannot keep up with natural language acquisition: finding large enough corpora is one of the most difficult problems.

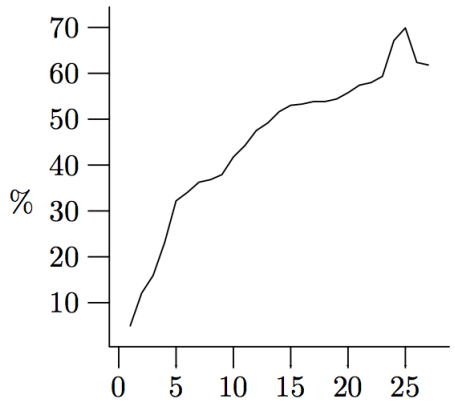

Figure 3: The recall values of likelihood ratio test.

6. References

A. Babarczy. 2002. *A Path from Broader to Narrower Grammars*. Ph.D. thesis, University of Edinburgh.

M. Bowerman. 1989. Learning a semantic system: What role do cognitive predispositions play? In R. L. Schiefelbusch and M. L. Rice, editors, *The Teachability of Language*. Paul H Brooks Publishing Co.

M. R. Brent. 1993. From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics*, 19:243–262.

C. J. van Rijsbergen. 1979. *Information retrieval*. Butterworths.

E. V. Clark. 1987. The principle of contrast: A constraint on language acquisition. In B. MacWhinney, editor, *Mechanisms of Language Acquisition*. Lawrence Erlbaum, Hillsdale, NJ.

D. Csendes, J. Csirik, and T. Gyimóthy. 2004. The Szeged corpus: A POS tagged and syntactically annotated Hungarian natural language corpus. In *Proceedings of TSD 2004*, volume 3206, Brno, Czech Republic.

K. Gábor and E. Héja. 2007. Clustering Hungarian verbs on the basis of complementation patterns. In *Proceedings of the ACL 2007 Student Research Workshop*, pages 91–96.

P. Halácsy, A. Kornai, L. Németh, A. Rung, I. Szakadát, and V. Trón. 2004. Creating open language resources for Hungarian. In *Proceedings of the 4th international conference on Language Resources and Evaluation (LREC2004)*.

P. Halácsy, A. Kornai, and Cs. Oravecz. 2007. Hunpos

– an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 209–212, Prague, Czech Republic. Association for Computational Linguistics.

D. Ienco, S. Villata, and Bosco C. 2008. Automatic extraction of subcategorization frames for Italian. In *Proceedings of the Sixth Language Resources and Evaluation (LREC 08)*, Marrakech, Morocco.

A. Korhonen, G. Gorrell, and D. McCarthy. 2000. Statistical filtering and subcategorization frame acquisition. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 199–206, Hong Kong.

A. Kornai, P. Rebrus, P. Vajda, P. Halácsy, A. Rung, and V. Trón. 2004. Általános célú morfológiai elemző kimeneti formalizmusa [output formalism of a general morphological analyser]. In Z. Alexin and D. Csendes, editors, *2nd Hungarian Computational Linguistics Conference*, pages 172–176, Szeged, Hungary.

A. Kornai, P. Halácsy, V. Nagy, Cs. Oravecz, V. Trón, and D. Varga. 2006. Web-based frequency dictionaries for medium density languages. In Adam Kilgarriff and Marco Baroni, editors, *Proceedings of the 2nd International Workshop on Web as Corpus*, pages 1–9.

Ch. Manning. 1993. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL '93)*, pages 235–242, Columbus, Oh.

M. Maragoudakis, K. Kermanidis, and G. Kokkinakis. 2000. Learning subcategorization frames from corpora: A case study for modern Greek. In *Proceedings of COMLEX 2000, Workshop on Computational Lexicography and Multimedia Dictionaries*, pages 19–22, Kato Achaia, Greece.

B. Sass. 2006. Extracting idiomatic Hungarian verb frames. In T. Salakoski, F. Ginter, S. Pyysalo, and T. Pahikkala, editors, *Advances in Natural Language Processing. 5th International Conference on NLP, FinTAL 2006*, pages 303–309, Turku, Finland.

S. Schulte im Walde. 2008. The induction of verb frames and verb classes from corpora. In A. Lüdeling and M. Kytö, editors, *Corpus Linguistics. An International Handbook*. Mouton de Gruyter.

V. Trón, P. Halácsy, P. Rebrus, A. Rung, P. Vajda, and E. Simon. 2006. Morphdb.hu: Hungarian lexical database and morphological grammar. In *Proceedings of 5th International Conference on Language Resources and Evaluation*, pages 1670–1673. ELRA.

D. Zeman and A. Sarkar. 2000. Automatic extraction of subcategorization frames for Czech. In *Proceedings of the International Conference on Computational Linguistics (COLING '00)*, pages 691–697.

Exploiting Comparable Corpora for Building and Expanding Terminological Resources

Fatiha Sadat
Université du Québec a Montréal
201 av. President Kennedy, Montréal, QC, H3X 2Y3, Canada
E-mail: sadat.fatiha@uqam.ca

Abstract

This paper seeks to present our interest in exploiting comparable corpora for building and expanding linguistic and terminological resources. Among these resources, our interest is focused on multilingual dictionaries, machine translation and ontologies.
Our past contribution in learning bilingual terminology from scarce resources was very attractive and positive, especially when applied to Cross-Language Information retrieval.
In this paper, we describe our past investigation in exploiting comparable corpora in order to translate and expand terms from source language to target language and possibly retrieve documents across languages. An extracted bilingual lexicon from comparable corpora provides a valuable resource to enrich existing bilingual dictionaries and thesauri. Also, a linear combination involving the extracted bilingual terminology from comparable corpora, readily available bilingual dictionaries and transliteration was proposed to Cross-Language Information Retrieval. An application on Japanese-English language pair of languages showed that the proposed combination yields better translations and an effectiveness of information retrieval could be achieved across languages.
Last, we describe the work in progress and other interesting applications of the automatic acquisition and production of language resources using large scale comparable corpora.

1. Introduction

Large text corpora represent a crucial resource for bilingual terminology acquisition and multilingual lexical resources enrichment. Moreover, in recent years non-aligned comparable corpora have been an object of studies and research related to natural language processing and information retrieval (Dagan and Itai 1994; Dejean et al. 2002; Diab and Finch 2000; Fung 2000; Koehn and Knight 2002; Nakagawa 2000; Peters and Picchi 1995; Rapp 1999; Shahzad and al. 2001; Tanaka and Iwasaki 1996; Daille and Morin, 2005, 2009), because of their availability and easy accessibility through the World Wide Web.

In this present paper, our goal consists in learning translation lexicons using scarce resources, i.e. readily available resources and possibly through the Internet. We are concerned by exploiting news articles as comparable corpora in order to translate terms in a source language to any specified target language. Our preliminary study is conducted on Japanese-English language pair using general-domain comparable corpora and could be extended to other languages and domains. Evaluations were conducted on Cross-Language Information Retrieval (CLIR) using a large-scale test collection NTCIR[1] for (Japanese, English) language pair. CLIR consists of retrieving documents written in one language using query terms in another language.

The remainder of the present paper is organized as follows: Section 2 presents an overview of the proposed approach for bilingual terminology acquisition from comparable corpora. Linear combination to dictionary-based translation and transliteration is presented in Section 3. Experiments and evaluations in CLIR are discussed in Sections 4. Section 5 concludes the present paper with work in progress and present and future extension.

[1] http://research.nii.ac.jp/ntcir/

2. An Overview of the Proposed Approach

Unlike parallel texts, which are clearly defined as translated texts, there is a wide variation of non-parallel-ness in monolingual data. It can be manifested in the topic, the domain, the authors, the time period, etc. Comparable corpora are collections of texts from pairs or multiples of languages, which can be contrasted because of their common features. We rely on such comparable corpora for the extraction of bilingual terminology in order to enrich existing bilingual dictionaries, thesauri and retrieve documents across different languages.

In the present study (Sadat et al., 2003; Sadat, 2004) , we follow the proposed model by (Dejean et al. 2002; Fung 2000; Rapp 1999). First, word frequencies, context word frequencies in surrounding positions (here three-words window) are estimated following statistics-based metrics. Context vectors for each term in the source language and the target language are constructed. We use the *log-likelihood ratio* (Dunning 1993) as expressed in equation (1):

$$\mathrm{LLR}(w_i, w_j) = K_{11}\log\frac{K_{11}N}{C_1R_1} + K_{12}\log\frac{K_{12}N}{C_1R_2} + K_{21}\log\frac{K_{21}N}{C_2R_1} + K_{22}\log\frac{K_{22}N}{C_2R_2} \quad (1)$$

where,
$C_1 = K_{11} + K_{12}$, $C_2 = K_{21} + K_{22}$,
$R_1 = K_{11} + K_{21}$, $R_2 = K_{12} + K_{22}$,

$N = K_{11} + K_{12} + K_{21} + K_{22}$,
K_{11} = frequency of common occurrences of word w_i and word w_j,
K_{12} = corpus frequency of word w_i - K_{11},
K_{21} = corpus frequency of word w_j - K_{11},
$K_{22} = N - K_{12} - K_{22}$.

Next, context vectors of the target words are translated using a preliminary seed lexicon. We consider all translation candidates, keeping the same context frequency value as the source term. This step requires a seed lexicon that will be enriched using the proposed bootstrapping approach of this paper.
Similarity vectors are constructed for each pair of source term and target term using the *cosine metrics* (Salton and McGill, 1983), as expressed in equation (2):

$$Similarity(w_i, w_j) = \frac{\sum_k v_{ik} v_{jk}}{\sqrt{\sum_k v_{ik}^2 \sum_k v_{jk}^2}} \quad (2)$$

where,
v_{ik} represents co-occurrence frequencies in context vectors of the source term w_i with term w_k. and v_{jk} represents co-occurrence frequencies in context vectors of the target term w_j with term w_k.
Therefore, similarity vectors are constructed to yield a probabilistic translation model $P_{comp}(t|s)$ for bilingual terminology extraction from comparable corpora.

3. Linear Combination

Combining different models has showed success in previous research (Dejean et al. 2002, Dejean et al. 2005). We propose a combined probabilistic translation model involving comparable corpora, readily available bilingual dictionaries as well as transliteration for the special phonetic or spelling representation of Japanese language, represented by the *Katakana* alphabet.
General-purpose dictionaries are basic source of translations and could be exploited for bilingual terminology extraction. The proposed dictionary-based translation model is derived directly from readily available bilingual dictionaries, by considering all translation candidates and their associated phrases, for each source entry.
Transliteration is the phonetic or spelling representation of one language using the alphabet of another language. The special phonetic alphabet (here Japanese *katakana*) to foreign words and loanwords requires *romanization* or transliteration (Knight and Graehl 1998). Japanese vocabulary is frequently imported from other languages, primarily (but not exclusively) from English. *Katakana*, the special phonetic alphabet is used to write down foreign words and loanwords, example names of persons and other terms.
Finally, translation alternatives are ranked according to the combined probability. A fixed number of top-ranked translation candidates are selected for each source term and misleading candidates are discarded.

The English word '*computer*' is transliterated in Japanese
katakana as 'コンピューター', as well '*engineer*' is transliterated as 'エンジニアー', and '*space shuttle*' is transliterated as 'スペースシャトル'. Named entities such as proper names of foreign (else than Japanese) persons, locations and organizations, are transliterated in Japanese. An example is '*Bill Clinton*' as named entities and transliterated in Japanese as 'ビルクリントン'. Therefore, the combined probabilistic model will involve distribution probabilities derived from the comparable corpora $P_{comp}(t|s)$, readily available bilingual dictionaries $P_{dict}(t|s)$ and the transliteration model $P_{translit}(t|s)$ as expressed in equation (3):

$$P(t|s) = \alpha_1 P_{comp}(t|s) + \alpha_2 P_{dict}(t|s) + \alpha_3 P_{translit}(t|s) \quad (3)$$

Parameters α_1 to α_3 are models dependant and represent the importance of each translation strategy, with $\sum_{i=1..3} \alpha_i = 1$.

4. Experiments and Evaluations

Experiments have been carried out to measure the improvement of our proposal on bilingual Japanese-English tasks in CLIR, i.e. Japanese queries to retrieve English documents.

4.1 Linguistic resources

A set of linguistic resources was used in these experiments and defined as follows:

- A collection of news articles from *Mainichi Newspapers* (1998-1999) for Japanese and *Mainichi Daily News* (1998-1999) for English are considered as comparable corpora, because of their common feature of the time period. Moreover, documents of *NTCIR-2* test collection were considered as comparable corpora in order to cope with special features of the test collection during evaluations.
- Morphological analyzers, *ChaSen* [2] version 2.2.9 (Matsumoto et al. 1997) for texts in Japanese and *OAK*[3] (Sekine 2001) for English texts were used in linguistic pre-processing.
- *EDR* (EDR 1996) and *EDICT* [4] bilingual Japanese-English dictionaries were used in translation.
- *KAKASI*[5], a language processing inverter and free software, available on the Internet was used in the transliteration process of Japanese terms written in

[2] http://chasen.aist-nara.ac.jp/

[3] http://nlp.cs.nyu.edu/oak/

[4] http://www.csse.monash.edu.au/~jwb/wwwjdic.htm

[5] http://kakasi.namazu.org/

katakana to English. Corrections on transliteration were completed manually by a native Japanese language speaker.

- *NTCIR-2* (Kando 2001), a large-scale test collection was used to evaluate the proposed strategies in CLIR.
- *SMART* information retrieval system (Salton 1971), which is based on vector model, was used to retrieve English documents.

4.2 Results and Discussion

Content words (nouns, verbs, adjectives, adverbs) were extracted from English and Japanese corpora. In addition, foreign words (mostly represented in katakana) were extracted from Japanese texts. Thus, context vectors were constructed for Japanese and English terms. Similarity vectors were constructed for Japanese-English pairs of terms.

We conducted experiments and evaluations on the monolingual and bilingual tasks of NTCIR test collection.

Topics 0101 to 0149 were considered and key terms contained in fields, title *<TITLE>*, description *<DESCRIPTION>* and concept *<CONCEPT>* were used to generate 49 queries in Japanese and English.

Results and performances of different translation models and their combination are described in Table 1. Evaluations were based on the average precision, differences in term of average precision of the monolingual counterpart and the improvement over the monolingual counterpart.

The combined dictionary-based and transliteration model *'DT'* showed 84.94% improvement of the monolingual retrieval *'ME'* while the comparable corpora-based model '*SCC*' showed a lower improvement in average precision compared to the monolingual retrieval and the combined dictionary-based and transliteration model *'DT'* with 52.81% of the monolingual retrieval.

The proposed combination of comparable corpora, bilingual dictionaries and transliteration '*DT&SCC'* showed the best performance in terms of average precision with 88.18% of the monolingual counterpart, +3.82% compared to the dictionary-based method and +66.97 compared to the comparable corpora model taken alone.

Translation Model	Avg. Precision	% Monolingual	% Difference (Improvement)		
ME	0.2683	100	-	-	-
DT	0.2279	84.94	-15.05	-	-
SCC	0.1417	52.81	-47.18	-37.82	-
DT&SCC	0.2366	88.18	-11.81	+3.82	+66.9

Table 1: Results and Evaluations on different translation models and their combination

5. Conclusion

We investigated the approach of extracting bilingual terminology from comparable corpora with an application on Japanese-English language pair. A combined model involving comparable corpora, readily available bilingual dictionaries and transliteration was found very efficient and could be used to enrich bilingual lexicons and thesauri. Most of the selected terms were considered as translation candidates or expansion terms in CLIR. Exploiting different translation models revealed to be effective.

These extracted terms could be very helpful in building a bilingual terminological resources and/or expanding an existing one. Moreover, this approach can be exploited efficiently in machine translation to cope with out-of-vocabulary words when using parallel corpora.

Comparable corpora are considered as very useful resources in many application of computational linguistics.

Ongoing research is focused on using and extending the proposed approach for machine translation and other applications of Cross-Language Information Retrieval.

We have a big interest in multilingual ontologies for under-resourced languages. We are interested in using comparable corpora for developing, extending and merging existing ontologies (exemple WordNet) for Arabic, English, French and Japanese.

6. References

Daille, B. Découverte et exploitation des corpus comparables pour l'accès à l'information multilingue (DECO). In Programme interdisciplinaire TCAN - Atelier de la plate-forme AFIA 2007, Grenoble, Grenoble, July 2007. AFIA 2007. (2007)

Daille, B., and Morin E.French-English Terminology Extraction from Comparable Corpora. *In Proceedings, 2nd International Joint Conference on Natural Language Processing(IJCLNP)*, Lecture Notes in Computer Sciences, vol. 3651, Springer, 707–718, Jeju Island, Korea. (2005)

Cancedda, N., Dejean, H., Gaussier, E., Renders, J., M., Vinokourov, A. Report on CLEF-2003 Experiments: Two Ways of Extracting Multilingual Resources from Corpora. CLEF 2003. (2003)

Dagan, I., Itai, I. Word Sense Disambiguation using a Second Language Monolingual Corpus. *Computational Linguistics* 20(4): 563-596. (1994)

Dejean, H., Gaussier, E., Sadat, F. An Approach based on Multilingual Thesauri and Model Combination for Bilingual Lexicon Extraction. In *Proceedings of COLING'02*, Taiwan, pp 218-224. (2002)

Dejean, H., Gaussier, E., Sadat, F. Automatic processing of multilingual medical terminology: applications to thesaurus enrichment and cross-language information retrieval. Artificial Intelligence in Medicine 33(2)-2005. (2005)

Diab, M., Finch, S. A Statistical Word-Level Translation Model for Comparable Corpora. In *Proceedings of the Conference on Content-based Multimedia Information Access RIAO*. (2000)

Dunning, T. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational linguistics* 19(1): 61-74. (1993)

EDR. Japan Electronic Dictionary Research Institute, Ltd. EDR electronic dictionary version 1.5 technical guide. *Technical report TR2-007, Japan Electronic Dictionary research Institute, Ltd*. (1996)

Fung, P. A Statistical View of Bilingual Lexicon Extraction: From Parallel Corpora to Non-Parallel Corpora. *In Jean Véronis, Ed. Parallel Text Processing*. (2000)

Gœuriot, L., Daille, B., and Morin, E. Compilation of specialized comparable corpus in French and Japanese. Proceedings, *ACL-IJCNLP workshop "Building and Using Comparable Corpora" (BUCC 2009)*, august 2009, Singapore. (2009)

Gœuriot, L., Morin, E., and Daille, B. Reconnaissance de critères de comparabilité dans un corpus multilingue spécialisé. Actes, *Sixième édition de la Conférence en Recherche d'Information et Applications (CORIA 2009)*. (2009)

Kando, N. Overview of the Second NTCIR Workshop. In Proceedings of the Second NTCIR Workshop on Research in Chinese and Japanese Text Retrieval and text Summarization, Tokyo. (2001)

Knight, K., Graehl, J. Machine Transliteration. Computational Linguistics 24 (4). (1998)

Koehn, P., Knight, K. Learning a Translation Lexicon from Monolingual Corpora. In *Proceedings of ACL-02 Workshop on Unsupervised Lexical Acquisition.* (2002)

Matsumoto, Y., Kitauchi, A., Yamashita, T., Imaichi, O., and Imamura, T. *Japanese morphological analysis system ChaSen manual.* Technical report NAIST-IS-TR97007, NAIST. (1997)

Morin, E., Daille, D., Takeuchi, K., and Kageura, K. Bilingual Terminology Mining -- Using Brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)* p. 664-671, Prague, Czech Republic, 2007. (2007)

Morin, E., and Daille, B. Comparabilité de corpus et fouille terminologique multilingue. *Traitement Automatique des Langues (TAL)*, 47(1):113-136, 2006. (2006)

Morin, E., Daille, B. Extraction de terminologies bilingues à partir de corpus comparables d'un domaine spécialisé. *Traitement Automatique des Langues (TAL)*, Lavoisier, 45(3), 103–122. (2004)

Nakagawa, H. Disambiguation of Lexical Translations Based on Bilingual Comparable Corpora. In *Proceedings of LREC2000, Workshop of Terminology Resources and Computation WTRC2000*, pp 33-38. (2000)

Peters, C., Picchi, E. Capturing the Comparable: A System for Querying Comparable Text Corpora. In *Proceedings of the Third International Conference on Statistical Analysis of Textual Data*, pp 255-262. (1995)

Rapp, R. Automatic Identification of Word Translations from Unrelated English and German Corpora. In *Proceedings of EACL'99*. (1999).

Sadat, F., Yoshikawa, M., Uemura, S. Learning bilingual translations from comparable corpora to cross-language information retrieval: hybrid statistics-based and linguistics-based approach. In *Proceedings of EACL'203, workshop on Information retrieval with Asian languages - Volume 11, Sapporo, Japan. Pages: 57 – 64. 2003.*

Sadat, F., Yoshikawa, M., and Uemura, S**.** Enhancing Cross-language Information Retrieval by an Automatic Acquisition of Bilingual Terminology from Comparable Corpora. *In Proceedings of the 26th Annual International* ACM SIGIR Conference on Research and Development in Information Retrieval SIGIR 2003. Toronto, Canada, Jul. 28-Aug. 1, 2003.

Sadat, F. Knowledge Acquisition from Collections of News Articles to Cross-language Information Retrieval. *In Proceedings of RIAO 2004 conference (Recherche d'Information Assisté par Ordinateur)*, Avignon, France, *pp. 504-513*, Apr. 26-28, 2004.

Salton, G. The SMART Retrieval System, Experiments in Automatic Documents Processing. *Prentice-Hall, Inc., Englewood Cliffs, NJ.* (1971)

Salton, G., McGill, J. *Introduction to Modern Information Retrieval.* New York, Mc Graw-Hill. (1983)

Sekine, S. *OAK System– Manual.* New York University. (2001)

Shahzad, I., Ohtake, K., Masuyama, S., Yamamoto, K. (1999) Identifying Translations of Compound Using Non-aligned Corpora. In *Proceedings of Workshop MAL*, pp 108-113.

Tanaka, K., Iwasaki, H. Extraction of Lexical Translations from Non-Aligned Corpora. In *Proceedings of COLING'96.* (1996)

Inducing Lexicons of Formality from Corpora

Julian Brooke, Tong Wang, Graeme Hirst

Department of Computer Science, University of Toronto
Toronto, ON, Canada M5S 3G4
jbrooke, tong, gh@cs.toronto.edu

Abstract

The spectrum of formality, in particular lexical formality, has been relatively unexplored compared to related work in sentiment lexicon induction (Turney and Littman, 2003). In this paper, we test in some detail several corpus-based methods for deriving real-valued formality lexicons, and evaluating our lexicons using relative formality judgments between word pairs. The results of our evaluation suggest that the problem is tractable but not trivial, and that we will need both larger corpora and more sophisticated methods to capture the full range of linguistic formality.

1. Introduction

The derivation of lexical resources for use in computational applications has been primarily focused on the semantic or denotational relationships among words, for instance the synonym and hyponym relationships encapsulated in a database like WordNet (Fellbaum, 1998). Largely missing from from popular resources like WordNet and the General Inquirer (Stone et al., 1966) is information about the formality of a word, which relates directly to the appropriateness of a word in a given context. The concept of formality has of course received a certain amount of interest in computational linguistics, for instance in studies of text generation (Hovy, 1990; Inkpen and Hirst, 2006). The lexical work on formality, however, generally assumes a static, discrete conception of formality. Theoretical and empirical work on genre and register (Leckie-Tarry, 1991; Biber, 1995; Heylighen and Dewaele, 1998) belies this idea; instead, linguistic formality and the dichotomies that underlie formality, e.g. spoken/written, interpersonal/abstract, contextual/context-independent, are generally conceived as dimensions, clines, or spectrums upon which particular genres may vary. Quantification of this spectrum, however, is rarely pursued beyond calculation of the easily countable surface features such as part of speech, providing broad metrics for text classification, but very little that can be applied to more subtle tasks such as word choice. In *Choose the Right Word* (Hayakawa, 1994), a manual intended to help writers select the best English word from among a group of near-synonyms, there is a clear assumption that the notion of a formality spectrum also applies at the lexical level; there, small differences between the formality of words are enumerated using relative, continuous language (i.e. *A is more formal than B* rather than *A is formal*).

In this work, we investigate methods for deriving a continuous spectrum of formality in the form of a formality lexicon. Our work is inspired and informed by the recent interest in sentiment lexicon acquisition that has formed a major part of work in Sentiment Analysis (Turney and Littman, 2003; Esuli and Sebastiani, 2006; Taboada and Voll, 2006; Rao and Ravichandra, 2009). We believe that construction of formality lexicons is a related but distinct problem, and so we will adapt methods used in the Sentiment research but also apply techniques which are distinct to the variations of formality. We predict that formality is a somewhat easier problem, due to the stronger co-occurrence relationships among formal words. One of the goals of this preliminary work is to show, however, that quantifying formality is far from trivial, particularly if the relationships among words are to be applied to tasks that require attention to linguistic detail, for instance word choice or phrase-level formality classification. More generally, we believe that a deeper understanding of formality may lead to applications that allow for capturing the variation of language in ways that avoid the pitfalls of domain-specificity, e.g. the need to train models for any possible location on the spectrum of register. Finally, one of our key goals is to develop methods for deriving lexical formality that are language-independent; the small-scale, corpus-based methods we investigate here are suitable for almost any language for which a varied corpus of a reasonable size is available.

2. Data and Resources

2.1. Word Lists

As the starting point for this work, we collected two lists of words, one formal and one informal, that we use both as seeds for our dictionary construction methods and as test sets for evaluation (our 'gold standard'). We assume that all slang terms are by their vary nature informal and so our 138 informal seeds[1] were pulled primarily from an online slang dictionary[2] (e.g. *wuss*, *grubby*) and also includes some contractions and interjections (e.g. *cuz*, *yikes*). The 105 formal seeds[3] were selected from a list of discourse markers (e.g. *moreover*, *hence*) and adverbs from a sentiment lexicon (e.g. *preposterously*, *inscrutably*); these sources were chosen to avoid words with overt topical content, and to ensure that there was some balance of emotional bias across formal and informal seed sets. The imbalance in the seed set counts (more informal than formal) is offset here by the fact that our formal seeds are much better represented in

This work is supported by the Natural Sciences and Engineering Research Council of Canada.

[1] See http://www.cs.toronto.edu/~jbrooke/informal_seeds.txt

[2] http://onlineslangdictionary.com/

[3] See http://www.cs.toronto.edu/~jbrooke/formal_seeds.txt

our primary corpus.

To allow for a more objective, fine-grained evaluation, we manually extracted a set of 399 pairs of near-synonyms[4] from *Choose the Right Word* (CTRW); all these pairs were either explicitly or implicitly compared for formality in the book. Implicit comparison included blanket statements like *this is the most formal of these words*; in those cases, and more generally, we avoided words appearing in more than one comparison (there are no duplicate words in our CTRW pair set), as well as multiword expressions and words whose formality is strongly ambiguous (i.e. word-sense dependent). An example of this last phenomenon is the word *cool*, which is used colloquially in the sense of *good* but more formally as in the sense of *cold*. Partly as a result of this polysemy, which we observe is more common among informal words, our pairs are clearly biased toward the formal end of the spectrum; although there are some informal comparisons, e.g. *bellyache/whine*, *wisecrack/joke*, more typical pairs include *determine/ascertain* and *hefty/ponderous*. Despite this imbalance, one obvious advantage of using near-synonyms in our evaluation metric is that factors other than linguistic formality (e.g. topic, opinion) are less likely to influence performance.

2.2. Corpora

Our primary corpus for the word co-occurrence methods presented here (section 3.3) is the Brown corpus (Francis and Kučera, 1982). Although extremely small by modern corpus standards, it has the advantage of being compiled explicitly to represent a range of American English genres (and, by extension, formalities). It includes four genres (reportage, formal documents, fiction, and miscellaneous) divided into 15 sub-genres; for our split-corpus method, we consider reportage and formal documents as formal. Its small size (approximately 1 million words in 499 documents) means that our results using it are likely to represent a lower bound rather than anything approaching optimal performance; nonetheless, we have found that it serves as a useful development set for selecting appropriate methods and testing various options. We note here that it contains at least one use of 53 (38%) of our informal seeds and 71 (67%) of our formal seeds. For our word count comparison methods (section 3.2) it is also useful to have a spoken corpus, representing the more informal end of the formality spectrum: for this, we use word counts for another publicly available corpus, the Switchboard (SW) corpus of American telephone conversations (Godfrey et al., 1992), which contains roughly 2400 conversations with over 2.6 million word tokens.

3. Methods

Each method described below derives a formality score (FS) in the range 1 to -1 for any word within its coverage, similar to the quantification of SentiWordNet (Esuli and Sebastiani, 2006). Since some methods do not have full coverage, in our evaluation we will also sometimes consider hybrid methods that back-off to a higher coverage (baseline) model; we do not, however, test more-complex hybrid systems (e.g. weighted sums) here.

[4] See http://www.cs.toronto.edu/~jbrooke/CTRWpairs.txt

3.1. Baselines

The most obvious baseline is based on word length, which is often used directly as an indicator of formality for applications like genre classification (Karlgren and Cutting, 1994). Given a shortest word of length n and a longest word of length m in some vocabulary V (the Brown corpus), we derive FS scores for any word based on this set by dividing up the formality scale into equal partitions; for a word w of length l, the formality score function, $FS(w)$, is given by:

$$FS(w) = -1 + 2\frac{l}{m-n}$$

A special exception is made for hyphenated terms, which can be extremely long in the case when an entire phrase is hyphenated, biasing the maximum word length: for those terms, we use the average length of constituent words rather than the total length. Though this metric works fairly well for English, we note that it might be problematic in a language with word agglutination (e.g. German) or without an alphabet (e.g. Chinese).

Another straightforward baseline is the assumption that Latinate prefixes and suffixes are indicators of formality in English (Kessler et al., 1997), i.e. informal words will not have Latinate affixes like *-ation* and *intra-*. Here, we simply assign words that have appear to have such a prefix or suffix an FS of 1, and all other words an FS of -1.

3.2. Frequency Methods

These methods derive FS based on word counts in corpora. Our first approach assumes a single corpus, where formal words are common and informal words are rare, or vice versa. To smooth out the Zipfian distribution, we use the rank of words as exponentials; for a corpus with R ranks, the FS for a word of rank r under the *formal is rare* assumption is given by:

$$FS(w) = -1 + 2\frac{e^{(r-1)}}{e^{(R-1)}}$$

Under the *informal is rare* assumption:

$$FS(w) = 1 - 2\frac{e^{(r-1)}}{e^{(R-1)}}$$

A more sophisticated method is to use two corpora that are known to vary with respect to formality and use the relative appearance of words in each corpus as the metric. If word appears n times in a (relatively) formal corpus and m times in an informal corpus (and one of m, n is not zero), we derive:

$$FS(w) = -1 + 2\frac{n}{m \times N + n}$$

Here, N is the ratio of the size (in tokens) of the informal corpus (IC) to the formal corpus (FC). We need the constant N so that an imbalance in the size of the corpora does not result in an equivalently skewed distribution of FS.

A hybrid method combines these two models by using the ratio of word counts in two corpora to define the center of the FS spectrum, but single corpus methods to define the edges. Formally, if m and n (word counts for the IC and FC, respectively) are both non-zero, then FS is given by:

$$FS(w) = -0.5 + \frac{n}{m \times N + n}$$

However, if n is zero, FS is given by:

$$FS(w) = -1 + 0.5\frac{e^{(r_{IC}-1)}}{e^{(R_{IC}-1)}}$$

where r_{IC} is the rank of the word in IC, and R_{IC} is the total number of ranks in IC. If m is zero, FS is given by:

$$FS(w) = 1 - 0.5\frac{e^{(r_{FC}-1)}}{e^{(R_{FC}-1)}}$$

where i is the rank of the word in IC, and R_{IC} is the total number of ranks in IC). This function is undefined in the case where m and n are both zero. Here we also consider the effect of lemmatization, treating various inflected forms as a single type.

3.3. Co-occurrence Methods

We test the co-occurrence methods used by Turney and Littman (2003) to derive Semantic Orientation (positive or negative word bias), with small modifications specific to our situation. The general idea is to derive an FS value for any given word by calculating the degree of association between it and the words in our seed sets. One such metric of association is Pointwise Mutual Information (PMI) (Church and Hanks, 1990); we derive probabilities using a word versus document matrix, with the FS of each word calculated as follows:

$$FS(w) = \frac{1}{N}(\sum_{f\in F}\frac{P(w\&f)}{P(w)P(f)} - \sum_{i\in I}\frac{P(w\&i)}{P(w)P(i)})$$

Here, F is the list of formal seeds, I is the list of informal seeds, and N is a normalization factor, either $argmax|FS'(w_F)|$ (for all w $FS'(w) > 0$) or $argmax|FS'(w_I)|$ (for all w, $FS'(w) < 0$), where $FS'(w)$ is the calculation before normalization; this last insures that the FS will be the range 1 to -1. $P(w\&f)$ is the probability (the count) of the word appearing with a particular formal seed in the same document.

The other method used by Turney and Littman, Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997), is a technique for extracting information from a large corpus of texts by (drastically) reducing the dimensionality of a word–passage matrix, i.e. a matrix where the row vectors correspond to the appearance or (weighted) frequency of words in a set of passages (the columns). The mathematical basis for this transformation is singular value decomposition[5]; for the details of the matrix transformations as relevant to this task, we refer the reader to the discussion in Turney and Littman (2003). The number of columns in the compacted matrix is given by the factor k, an important variable in any application of LSA, and one that is best determined by trial and error. Another factor is the size of a passage, which could be as large as a full document or as small as a sentence; here, we consider documents and paragraphs as possible passages[6]. A third variable that we investigated is the weighting of values in the original matrix; Turney and Littman, for instance, used *tf·idf* (term frequency times inverse document frequency), however it was not clear that this was appropriate for our task, and so we tested various possible options (binary, *tf*, *idf*, and *td·idf*). Again, we consider the effect of lemmatization.

Once a k-dimensional vector for each word appearing in a corpus is derived using LSA, a standard method is to use the cosine of the angle between a word and sets of seed words to identify how similar the distribution of the word is to the seeds. In our case, FS calculated as:

$$FS(w) = \frac{1}{N}(\sum_{f\in F}\cos(\theta(w,f)) - \sum_{i\in I}\cos(\theta(w,i)))$$

Again, F and I are the formal and informal seed sets, and N is a normalization factor, calculated in the same way as with PMI, above.

Another method that is available to us, due to the relatively large size of our seed sets, is derivation of FS by means of regression, using machine learning algorithms. We speculate that this might be preferable to the cosine method since the irrelevant dimensions might be discarded from the model, whereas in the cosine calculations these dimensions would show up as noise. To investigate the effectiveness of this approach, we tested various regression algorithms included in the WEKA software suite (Witten and Frank, 2005); below, we present results for two, linear regression and Gaussian processes, which preformed well according to the r^2 value with 10-fold cross-validation; for both we used the default settings for WEKA (version 3.6.2), which for Gaussian processes entails a classifier with an RBF kernel. Training was carried out using the k-dimensional vectors of our formal and informal seeds; for the purposes of training the former were assigned a value of 1, the latter -1. Since the model applied to new data could potentially fall outside that range, appropriate normalization of the output (dividing by the most extreme FS values) is also necessary in this case.

4. Evaluation

We evaluate our lexicon dictionary methods using the gold standard judgments from the seed sets and CTRW word pairs. To differentiate the two, we continue to use the term *seed* for the former; in this context, however, these 'seed sets' are being used as a test set. For computation of the co-occurance-based FS of a word that is part of our seed set, we apply *leave-one-out* cross-validation, removing that word from list of seeds for the purposes of calculating cosine difference from the seeds or when training a model to predict its FS value. The coverage (Cov.) is the percentage of words in the set which appear in the induced dictionary. The class-based accuracy (C-Acc.) is the percentage of words which are correctly classified as formal (FS > 0) or informal (FS < 0). The pair-based accuracy (P-Acc.) is the result of exhaustively pairing words in the

[5] We use the Divisi Python implementation of SVD, http://divisi.media.mit.edu; our vectors are taken from 'weighted U' matrix after SVD is applied and all but the top k singular values are removed.

[6] Preliminary testing with sentences suggested that the resulting matrices were far too sparse to be useful, we omit those results here.

two seed sets and testing their relative formality; that is, for all $w_i \in I$ and $w_f \in F$, the percentage of w_i/w_f pairs where $FS(w_i) < FS(w_f)$. The average FS difference (FS-Dif.) is just $FS(w_i) - FS(w_f)$ for each of the w_i/w_f pairs created as above; we wish to maximize this number on the basis that our seeds represent relatively extreme examples of the formality spectrum. For the CTRW pairs there are only two metrics, the coverage and the pair-based accuracy; since the CTRW pairs represent relative formality of varying degrees, it is not possible to calculate a class-based accuracy and there is no guarantee that the average distance should be maximized.

5. Results

The results of evaluation for all the various methods are shown in Table 1; the numbers in parentheses below indicate the corresponding line of the table. In the first section of the table, the baseline provided by the word length (1) is quite high, particularly for seed set pairwise accuracy, indicating that nearly all the informal seed words are shorter than the formal seed words. Word length is not as effective with the fine-grained differences, however, and the class-based accuracy is low, as many formal seeds are incorrectly labeled as informal using our linear method.[7] It is clear from the class-based accuracy score that Latinate suffixes and prefixes (2) are indicative of formality; they do not, however, provide information that allows for relative, more fine-grained distinctions. The advantage of these methods, of course, is their coverage.

The first two results in the second part of Table 1 (3–4) show that neither assumption (i.e. that formal words are rare or that informal words are rare) is particularly successful, though they fail in different ways that are indicative of the formality make-up of the corpus and the test sets. Since the Brown corpus is a corpus of published written texts, and therefore more formal, the *informal is rare* hypothesis (3) is a better one for the extreme seed sets; however, in the CTRW test sets, which is more indicative of the formal end of the spectrum, this assumption fails spectacularly, with the model performing much worse than chance. The opposite is true for the *formal is rare* model (4), since it makes opposite predictions. Neither is directly useful for the task as a whole.

Much better is the word ratio model using the Brown corpus as the formal dictionary and the Switchboard corpus as the informal dictionary (5); although the coverage is quite low, the score for pairwise accuracy in the CTRW set is the highest in Table 1, and the scores for the seed test are also quite good. The hybrid model, with the ratio model converting the middle of the spectrum and the *rare* models applied at either end (6), provides us with the best class-based accuracy in the table, and comparable performance among CTRW pairs with a 20% increase in coverage. A hybrid model that splits the Brown corpus into two halves (7), i.e. the relatively formal genres of reportage and formal documents and the relatively informal genres of fiction and miscellaneous, does not, however, perform nearly as well, suggesting that very distinct corpora are required for this method to be useful. The effects of lemmatization (8) are harder to interpret; the drop for seed words is marked, but there is a modest increase for CTRW, and a small boost in coverage. In general, this suggests that the inflectional differences among words might be somewhat indicative of formality, and should not necessarily be disregarded. Finally, the use of a word length backoff (9) provides superior performance with respect to the seed sets, but is slightly worse than the word length baseline in the CTRW set.

[7] Switching to logarithmic FS function for word length would likely improve the class-based accuracy, though fine-tuning this function would take us beyond a simple baseline, and have no effect on the pairwise accuracy.

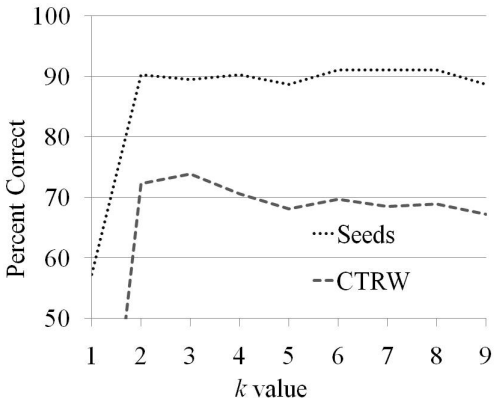

Figure 1: Seed class-based accuracy and CTRW pairwise accuracy, LSA cosine method for various k, $1 \leq k < 10$

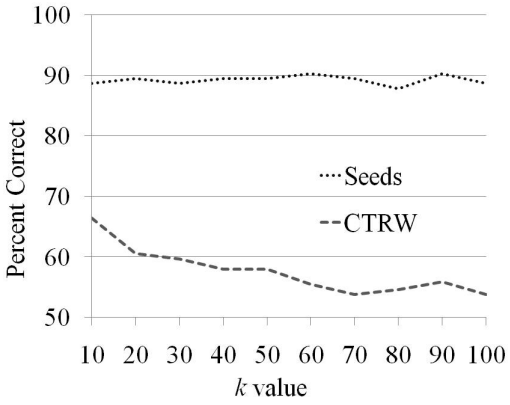

Figure 2: Seed class-based accuracy and CTRW pairwise accuracy, LSA cosine method for various k, $10 \leq k \leq 100$

The co-occurrence results are presented in the third part of Table 1. The PMI results (10) are quite promising, given the simple nature of the calculation, though the LSA results (11–14) are better, particularly when the optimal value of k is used (14). To find that value, we tested all values between 1 and 10, and at intervals of 10 thereafter; graphs showing the class-based accuracy for seeds and pairwise accuracy for the CTRW set are presented in Figures 1 and 2.

For the CTRW set, the performance peaks at $k = 3$; beyond $k = 3$, the overall trend is down, though there are small jumps at particular values of k, and we often see corresponding fluctuations in seed set performance, even though the overall picture there is much flatter. We posit

Dictionary construction method	Seed set				CTRW set	
	Cov.	C-Acc.	P-Acc.	FS-Dif.	Cov.	P-Acc.
Baseline methods						
(1) Word length	100	74.9	91.8	0.49	100	63.7
(2) Latinate affixes	100	74.5	46.3	0.86	100	32.6
Word count methods						
(3) Word Counts, Brown, informal is rare,	51	63.7	68.3	0.45	59	18.5
(4) Word Counts, Brown, formal is rare	51	36.3	19.5	−0.45	59	55.0
(5) Ratio, Brown and Switchboard	38	81.5	85.7	0.75	36	78.2
(6) Hybrid, Brown and Switchboard	58	90.8	89.4	0.81	56	74.3
(7) Hybrid, split Brown	51	51.6	70.0	0.49	60	38.2
(8) Hybrid, Brown and Switchboard, lemma	63	78.6	79.1	0.51	62	77.0
(9) Hybrid, Brown and Switchboard, WL backoff	100	87.7	92.3	0.72	100	61.2
Co-occurence methods						
(10) PMI, Brown	51	80.6	84.4	0.33	60	73.2
(11) LSA (k=100), cosine (Brown, binary, document)	51	88.7	96.1	0.74	60	53.8
(12) LSA (k=10), cosine	51	88.7	95.0	1.00	60	66.4
(13) LSA (k=3), cosine	51	89.5	94.5	1.07	60	73.9
(14) LSA (k=3), cosine, WL backoff	100	88.9	95.1	0.94	100	62.2
(15) LSA (k=3), cosine, lemma	51	88.5	94.4	1.02	60	70.5
(16) LSA (k=100), cosine, paragraph	51	83.1	96.6	0.65	60	53.8
(17) LSA (k=10), cosine, paragraph	51	83.1	95.0	0.86	60	61.8
(18) LSA (k=3), cosine, paragraph	51	83.1	91.7	0.86	60	73.5
(19) LSA (k=3), *tf*, cosine	51	66.1	74.9	49.2	60	49.2
(20) LSA (k=3), *idf*, cosine	51	55.6	57.7	0.02	60	52.5
(21) LSA (k=3), *td·idf*, cosine	51	54.8	39.7	−0.07	60	52.5
(22) LSA (k=100), Gaussian	51	71.8	83.8	0.42	60	38.2
(23) LSA (k=10), Gaussian	51	81.5	92.3	0.45	60	56.3
(24) LSA (k=3), Gaussian	51	87.1	92,7	0.39	60	56.7
(25) LSA (k=100), linear	51	58.9	57.6	0.04	60	53.4
(26) LSA (k=10), linear	51	79.0	88.9	0.12	60	58.4
(27) LSA (k=3), linear	51	75.8	86.8	0.14	60	61.8

Table 1: Seed coverage (%), class-based accuracy (%), pairwise accuracy (%), average FS difference, CTRW coverage (%) and pairwise accuracy (%) for various FS dictionaries

that the more fine-grained CTRW set is much more sensitive to the noise that comes with the increase in dimensionality; clearly, the second dimension (the one that is 'added' at $k = 2$) is the strongest indicator of formality, and though other dimensions (e.g. $k = 3, 6$) also provide information that boost performance. More generally, however, the addition of dimensions is a losing proposition, as the best dimensions for detecting formality are among the first discovered by using the LSA method, and beyond that the noise outweighs the relevant information. The results with a word length backoff suggest that overall the LSA method is slightly better than the hybrid word-count method, though the differences are not significant.

Looking at the options for LSA, lemmatization (15) has a small but consistently negative effect. More notable is the drop in performance when paragraphs rather than documents are taken as the unit in our word–passage matrix (16-18), suggesting that a *one formality per document* assumption is a relatively good one; the pairwise accuracy in the seed sets, though, is consistently high. With respect to weights, our original intuition was that a binary feature for appearance in a document was the best way to approach the construction of a word-document matrix; intuitively, there does not seem to be useful information that can be gleaned from the number of appearances of a formal or informal word in a document, nor should a word be weighted solely based on its rarity in a corpus. Indeed, our results (19–21) confirm this; applying *td·idf* or either of its component results in a major drop in performance across the board.

Finally, we look at the results using machine learning regression methods rather than cosine distance to derive FS (22–27). Neither of the algorithms perform well on the CTRW set, with the Gaussian Processes method (22–24) particularly poor, despite its relative sophistication; one explanation is that it tries to maximize the extreme cases, failing on the more-subtle word distinctions. The performance differences related to increases in k (22, 25) are consistent with cosine but more marked, revealing themselves in all three accuracy measures, though with a great deal more variation across the methods. Regression might prove to be more effective with more-numerous and more-nuanced training examples (for instance, including seed words that represent the middle of the spectrum).

One gratifying result is that, despite particular inconsistencies, the four performance metrics used here show clear correlation; for instance, even when the seed accuracies are

flat, increases in *k* are associated with both a drop in CRTW accuracy and a drop in the average FS difference between seed words. Thus, we can be confident that the performance differences among our models are robust, reflecting variation across the full spectrum of formality.

6. Conclusions and Future Work

Though preliminary, the work we have presented in this paper suggests that quantifying formality is a tractable but not trivial problem. Surprisingly, despite significant variation in the underlying features from which they are derived, several of the models investigated here reached an impressive accuracy in distinguishing extreme differences in formality, using information derived from a small yet diverse corpus. Less encouraging, however, is the performance of these same methods in identifying more-subtle variations among near-synonyms; at present, our guess based on the word count and co-occurrence is no better than one based simply on word length.

The next step in this project will involve an expansion of our data. There are a number of larger publicly available corpora that could be applied to our problem, for instance the British National Corpus (Burnard, 2000); informal testing suggests that word count information from the BNC will easily boost our word-count performance well beyond the baselines provided by word length. Blogs are a natural, inexhaustible source of information on register variation, though there are potential pitfalls and challenges related to using large amounts of web data, in particular the fact that LSA, our most promising method, does not scale up well (Turney and Littman, 2003).

With respect to refining our methods, one way forward is to see how the information represented by these various methods can be integrated to improve performance, i.e. with some kind of meta-classifier. There is certainly room for the methods to inform each other, since agreement for our best word count classifier and best co-occurrence classifier in the CTRW test set is a mere 66.3%, almost 10% below the accuracy in both cases; agreement on the seed sets is much higher, of course, but still below 90% for both metrics. One difficulty here is the lack of reliable training data, and one option we are exploring is the use of semi-automated methods to derive larger, more objective seed sets. A related idea is to use, for instance, word count or PMI FS as a starting point, and then use the LSA co-occurence information to iteratively refine those scores until convergence. In short, the methods described here just represent a basic toolbox for the continued exploration of the formality spectrum, moving beyond English-specific approaches to those that can be applied in any language.

7. References

Douglas Biber. 1995. *Dimensions of Register Variation: A cross-linguistic comparison*. Cambridge University Press.

Lou Burnard. 2000. User reference guide for British National Corpus. Technical report, Oxford University.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguististics*, 16(1):22–29.

Andrea Esuli and Fabrizio Sebastiani. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Internation Conference on Language Resources and Evaluation(LREC)*, Genova, Italy.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Nelson Francis and Henry Kučera. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin, Boston.

J.J. Godfrey, E.C. Holliman, and J. McDaniel. 1992. Switchboard: telephone speech corpus for research and development. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1:517–520.

S.I. Hayakawa, editor. 1994. *Choose the Right Word*. HarperCollins Publishers, second edition. Revised by Eugene Ehrlich.

Francis Heylighen and Jean-Marc Dewaele. 1998. Variation in the contextuality of language: An empirical measure. In *Context in Context, Special issue of Foundations of Science*, pages 293–340.

Eduard H. Hovy. 1990. Pragmatics and natural language generation. *Artificial Intelligence*, 43:153–197.

Diana Inkpen and Graeme Hirst. 2006. Building and using a lexical knowledge base of near-synonym differences. *Computational Linguistics*, 32(2):223–262.

Jussi Karlgren and Douglas Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th Conference on Computational Linguistics*, pages 1071–1075.

Brett Kessler, Geoffrey Nunberg, and Hinrich Schütze. 1997. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 32–38.

Thomas K. Landauer and Susan Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

Helen Leckie-Tarry. 1991. *Language Context: a functional linguistic theory of register*. Pinter.

Delip Rao and Deepak Ravichandra. 2009. Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Lingusitics*, Athens, Greece.

Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilivie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

Maite Taboada and Kimberly Voll. 2006. Methods for creating semantic orientation dictionaries. In *Proceedings of the 5th Internation Conference on Language Resources and Evaluation (LREC)*, Genova, Italy.

Peter Turney and Michael Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21:315–346.

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco.

Semantic resource extraction from Wikipedia category lattice

Olivier Collin, Benoît Gaillard, Jean-Léon Bouraoui
Orange Labs
Avenue Pierre Marzin, 22300 Lannion

Thomas Girault
Université de Rennes I
2, rue du Thabor 35065 Rennes

E-mail: olivier.collin@orange-ftgroup.com, benoit.gaillard@orange-ftgroup.com,
jeanleon.bouraoui@orange-ftgroup.com,toma.girault@gmail.com

Abstract

This work is closely related to the domain of automatic acquisition of semantic resources exploiting Wikipedia data. More precisely, we exploit the graph of parent categories linked to each Wikipedia page to perform a hierarchical parent categories extraction, semantically and thematically related. This extraction is the result of a shortest path length computation applied to the global lattice of Wikipedia categories. So, each page can be indexed by its first level categories, and in addition within their parent categories. This resource has been used for two kinds of applications. The first one concerns semantic query expansion for a multimedia search engine. The second one is a query translator for a multimedia search engine. This last work has been performed by using English lattice of categories and Wikipedia translation tables.

1. Introduction

This work is closely related to the very large area of lexical semantic resources constitution. The aim is to provide featuring labels for each lexical entry, these labels being hierarchically organised within a taxonomy or a lattice. This representation should lead to a generalisation of the input lexical space (hyperonyms), but also to homonym differentiation and synonym clustering. This point of view is usually shared by the linguistic community which targets a precise and exhaustive modelling of lexical entries. An alternative representation is a vector space approach. Lexical entries are modelled by a vector of neighbour words counts, these neighbours being extracted in the same document within a local window or not, and by using linguistic processing or not. This way, standard vectorial or statistical techniques can project each entry on a "semantic distributed subspace" (Latent Semantic Analysis for example) or within clusters of semantically related entries (K-means for example).

A similarity measure is usually associated to these representations so that one can express a kind of proximity between lexical entries. Then, this measure enables semantic expansion treatment (semantic proximity) or ambiguity resolution (semantic differentiation). This paper proposes an alternative representation space: a subset of the Wikipedia category lattice.

Since Wikipedia creation, many authors (Medelyan and al., 2008), (Suchanek and al., 2008), (Mihalcea, 2007), (Ponzetto and al., 2007), (Zesch and al., 2007), (Strube and al, 2006), have explored means of exploiting Wikipedia data to make a usable semantic resource. This paper is closely related to these preceding works: we automatically extract, without any human help, a sub-lattice from Wikipedia category lattice. The sub-lattice structure is linguistically imperfect but shows, in many cases, relevant hyperonymic relations and thematic categories. We have evaluated this resource relevancy on two use-cases: a semantic expansion task and a query translation task where obviously disambiguation is required.

2. Resource extraction

2.1 The Wikipedia categories lattice

Each Wikipedia page is indexed by a set of visible parent categories, which can be clicked in at the bottom of each page. So, parent french categories for "*Tom Cruise*" page are: *Acteur américain, Producteur américain, Naissance dans l'État de New York, Naissance en 1962, Personnalité américaine d'origine allemande, Personnalité américaine d'origine britannique, Personnalité américaine d'origine irlandaise, Scientologie.* These categories usually express one or more semantic roles. Each of these bottom page categories is also a Wikipedia page which has parent categories. This hierarchy of categories is not a strictly build taxonomy since categories and hierarchical links are freely added by various contributors. These contributions constitute a part of Wikipedia richness as a kind of *folksonomy*,, however semantic relations are difficult to extract from this constellation (Guégan 2006), (Strube 2006). This space of categories rather constitutes a graph oriented towards a set of a unique parent category "*Article*", so, it rather constitutes a lattice. This lattice is linguistically organised, usually each parent category is generalising each child category following a hyperonymic or thematic axis. For the moment, we can't separate these two axes. In addition, for each category, several

generalisations act in parallel. For example, in the case of "*Tom Cruise*" parent category *Acteur_américain*, a generalisation direction is: *Artiste_américain>Art_aux_Etats-Unis>Art_par_pays> Art>Article.* (son > parent).

2.2 Lattice fabrication

Raw data, page to category links and category to category links, come from two SQL tables[1]downloaded from the Wikipedia french resource site[2]. Suitable joints on these tables allow a direct relation between each Wikipedia page or category and its parents. This is a flat representation of the overall lattice which virtually enables us to list all the paths between articles and the terminal category "*article*". The page/category links are straight-forward, the combinatorial part of the lattice is mainly related to the category/category links sub-lattice. For computational purpose, we have separated page/category links from category/category links. Finally, for French data, we get 873 468 pages linked to 3 770 343 parent categories (4.31 category per page in average), and 119 492 categories linked by 244 817 edges (2.04 parent category per category on average). Our goal was to expand the first level of parent category pages like *Acteur américain* or *Producteur américain* for "*Tom Cruise*" example. So we only processed the category/category sub-lattice.

However, even if the connectivity of this graph is not so high and shows "small world" properties (Guegan, 2006), the quantity of such paths is too great (dozens or even hundreds of paths for each category) to be used without pre-processing. In addition, the flat representation of the lattice doesn't match with our navigational needs. So, we have used NetworkX package[3], which allows us to upload the flat representation tables up to memory. This package also provides many useful functions for a quick navigation in a graph. The overall lattice of categories (119 492 nodes and 244 817 edges) has been uploaded in memory and we have been able to test several quick search algorithms

So, the main challenge of our work has consisted in making a relevant selection among all paths for providing useful semantic data, especially for disambiguation purpose.

2.3 Sub-lattice extraction

The sub-lattice extraction is based on a strong assumption: the relevant information is carried by the shortest paths that link each of the pages to the terminal category. In fact, after some testing we realized that paths linking to the "*Article*"category were less relevant than the paths reaching the set of categories pointed to by the top level category page *Wikipedia:Catégorie*[4]. This set contains 150 pseudo terminal categories such as (in French): *"Mouvement culturel", "Art contemporain", "Artisanat","Design", "Art par pays", "Rayonnement culturel", "Artiste"*… Given the overall lattice and a shortest path calculus provided by NetworkX package, the filtering process is:

For each page
    For each parent category page
        Select the shortest paths that reach
            all terminal categories

We kept all the shortest paths of the same length which can occur for one category. For a same page, we don't keep paths of length greater than 8 and we only keep the 15 shortest paths. We have filtered a few initial categories like date or place of birth brought noisy paths. Here are results obtained for the two different French pages related to *"Avocat"* (fruit versus occupation):

Avocat_(fruit) (only one path)
1-Fruit_alimentaire>Plante_alimentaire> Plante_utile>Agriculture

Avocat_(métier)(two paths)
1-Métier_du_droit>Droit
2-Personnalité_du_droit>Droit

We observe paths that reach a global thematic category through hyperonymic relations. In this example, filtered data provides quite a good result from a linguistic point of view and shows obvious disambiguation possibilities. In other cases we get more noisy paths, but they are still relevant. The following paths form a sub-lattice related to *"Tom Cruise"* page:

acteur_américain>acteur_par_nationalité>acteur>personnalité>médias
acteur_américain>artiste_américain>art_aux_tats-Unis>art_par_pays>art
acteur_américain>artiste_américain>artiste_par_pays>artiste>personnalité>art
producteur_américain>cinéma_américain>cinéma_aux_états-Unis>cinéma_par_pays>art_par_pays>art
producteur_américain>producteur_de_cinéma_par_nationalité>producteur_de_cinéma>producteur>artiste>personnalité>art
scientologie>groupement_spirituel>spiritualité_autres>spiritualité
scientologie>groupement_spirituel>petit_mouvement_religieux>religion>spiritualité
portail:cinéma>portail:art>portail:culture
portail:états-unis>portail:Amérique>portail:géographie>portail_du_domaine_géographique

The sub-lattice attached to each page is not always very large, especially for French pages and these data are not well formed on a strict linguistic point of view. However, they give a good trade-off between quantity and quality of

[1] *Frwiki-latest-page.sql and frwiki-categorylinks.sql*
[2] http://download.wikimedia.org/frwiki/latest/.
[3] http://networkx.lanl.gov/
[4] http://fr.wikipedia.org/wiki/Wikipédia:Catégories

features. For application purpose, the hierarchy can be broken and these data can be used as a useful "bag of categories". They can help us to treat a disambiguation task by using classical constraints based on hyperonyms and thematic classes: *fruit, agriculture / personnalité, droit*.

A similar job has been done for English Wikipedia pages and categories[5]. The lattice of English categories is bigger than the French one (524 313 nodes and 1 206 219 edges) but NetworkX allows us to get a memory mapping and navigational functions still work quite well. We have applied the same filtering process and created an English resource. However, we didn't use pseudo terminal categories which are not relevant for English data. In addition, many administrative categories are polluting our representation space. For now, we have not filtered all these noisy categories but this first level of English resource is already useful.

These semantic resources (French, English) have been used and evaluated on two different tasks. In each case, the goal was to increase the recall or the relevance of a proprietary multimedia search engine. The first task consists in a kind of query expansion. We have re-structured our French semantic space with a concept lattice technique. The result enables to generate new queries close to the initial user query. The second task is a query translation task for cross lingual information retrieval (French to English). Our English semantic resource enables us to perform a choice between the different translation hypotheses by using a cosine measure in an associated vector space.

3. Query expansion task

The issue of query expansion is to add to the initial query some words that are "similar" to it, or even to use them to replace it. The goal is to give the user more relevant documents in regard to his query, even if they don't match with the initial terms. For example, for the query "car", the search engine will also retrieve documents that contain "automobile". A similar application is the content recommendation. It consists in proposing to the user some documents that do not match directly his query, but that should nonetheless interest him. In both cases, the aim is to model the similarity of the proposed terms. To overcome this problem, we use the resource described in section 2. A query corresponds to the name of a Wikipedia page. A concept lattice made from the resource allows us to extract relevant terms that are similar to the query.

3.1 Resources

Our resource enables us to index the Wikipedia pages, not only from their parent categories but also directly from the categories that take part in the associated sub-lattice. Thus, we can directly get all the pages that are indexed with *Acteur américain* but also all of the "acteurs" or the "personnalités". This is a first application of our work, and for now Wikipedia does not propose this level of indexation on all the pages. In order to restrict the quantity of data to process for this study, we selected the subset of Wikipedia pages indexed with the *"informatique"* topic (25 140 pages). The paths of all the categories associated to a page have been changed into one single "bag of categories". Finally, each page is represented by a vector of categories that contains all its parent categories that reach the terminal category (hyperonyms, topics). Though, this representation loses the initial hierarchy, it allows us to use some standard techniques of classification or "data-mining", that rely on vectors of features. However, the combination of specific categories such as *"Matériel_informatique"* and generic categories such as *"Informatique"* is still a very structuring information. For example, a part of the vector attached to *"Disque dur multimédia"* contains, among others:

Matériel_audio-vidéo, Audiovisuel, Médias, électronique, Multimédia,Informatique,Stockage_informatique,Matériel_informatique,Techniques_et_sciences_appliquées,Stockage_informatique, Industrie, économie…

The similarity between the pages is then computed from this representation.

3.2 Method overview

We generated a concept lattice from the data previously described by using an implementation from Girault (Girault, 2008). The concept lattice is made of "formal concepts". In this formalism, each formal concept is described by an extension and an intension. The extension is an enumeration of the set of the members of the same category. The intension is the set of the properties shared by the member of this category. In our work, a formal concept extension is a set of page names. The intension corresponds to the categories names, which are the items shared by the vector attached to the extension pages. That means that the names of the categories are used as features that will define the common points between the page names. We obtained a lattice made of 293 636 formal concepts that described the computer science domain. In our framework, an instance of a formal concept is:

Extension: *['Ethernet', 'Segment_de_réseau']*

Intension:*['électronique','Informatique','Télécommunications','Portail:Science','Protocole_réseau','Portail:Informatique','Normes_et_standards_informatiques','Portail:Technologie','Composant_électronique','Protocole_de_técommunication','Normalisation_des_técommunications','Techniques_et_sciences_appliquées','Matériel_informatique','Connectique']*

We obtained a pool of pages that share common categories. In this first study, the similarity of a page is defined by taking into account formal concepts which corresponds to the following criteria:

- The extension includes strictly 2 items; one of these items is the considered page;
- The intension includes at least 8 items.

This choice allows to link two pages that share more than eight categories. Our previous example respects this criterion; thus the page *"Segment_de_réseau"* is similar to

[5] http://download.wikimedia.org/enwiki/latest

the page *"Ethernet"* since they share 14 categories. The preliminary results that follow have been produced according to this computation.

3.3 Results

The first results concern query expansion. We chose as an example the query *"Ethernet"*. Within the relevant [6] formal concepts, this term is included in the extensions with: *Chiffreur IP, RS-232, IEEE 802.3, Protocole réseau, Informatique, Réseau informatique, Matériel informatique, IEEE 802, Segment de réseau, Architecture informatique, Carrier Sense Multiple Access with Collision Detection, Medium Attachment Unit.* All these terms have a neighbourhood link with the initial query. By the way, this link sometimes corresponds to some semantic link: notably synonymy (*IEEE 802.3*) and hyponymy (*Protocole réseau*). Most of the other terms are used in the context of "*Ethernet*", such as *Segment de réseau*, or *Carrier Sense Multiple Access with Collision …* Some of these expansions can be too specific in regard to the initial query, and add noise in the retrieved documents. A way to carry out this problem is to filter the proposed terms with the application index, as done in (Gaillard and al. 2010); thus, only the terms actually existing in the available documents are used. The following results use the same strategy but are obtained from queries about products. The applicative framework then becomes a recommendation system: the obtained data allows to the user, from an initial query, to be proposed other products likely to interest him. An emerging and promising feature of our results is their structuring: it allows to sort them according to several thematic dimensions. The figure 1 displays the thematic context of the query *"Super_Mario_Bros"*, as well as the associated directions of thematic associations.

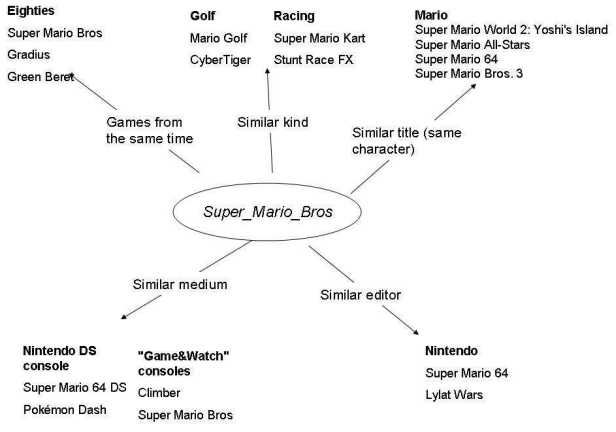

Figure 1: Game recommendations according to different thematics

Items of the extension *['Mario_Golf', 'CyberTiger']*, are linked to the intension *['Informatique', 'Projet_jeu_vidéo','Golf','Jeu_Nintendo_64','Jeu_vidéo'," Application_de_l'informatique","Jeu_vidéo_sorti_en_1999','Jeu_vidéo_de_golf','Sport_individuel','Techniques_et_sciences_appliquées','Sport','Audiovisuel','Projet:Jeu_vidéo', 'Médias']*. We plan to automatically sort the extensions thanks to some items in their intension (for instance, "Golf" is explicitly referred to in the intension above). These first results, very promising, show that Wikipedia data constitutes a first resource for the expansion and recommendation techniques. This work is confirmed by other works that use this encyclopedia for this objective (Tien-Chen and al. 2007), (Peng and al. 2008).

4. Query translation task

4.1 Resources

In addition to our English semantic resources, we have constituted a bilingual (French/English) dictionary from the translation table for French pages[7]. This table lists multilingual links from French articles to their equivalents in various languages of Wikipedia, provided they exist. Only French to English links have been kept. A joint between the table of French articles [8] and the translation table enabled us to get a direct relation between French pages and English pages. We ended up with a table that directly associates titles with their various translations: *Avocat (fruit)/Avocado or Avocat (métier)/Lawyer*, for example. This translation table is comparable to a bilingual dictionary having 540 920 entries. Its specificity is that it contains many named entities and phrases, such as: *Avocat du diable/ Devil's advocate; L'Avocat du diable (film)/Guilty as Sin*. This bilingual dictionary can therefore be used directly but offers no solution to make a choice among the various translation alternatives.

4.2 Method overview

First of all, queries are segmented in lexical units which can be simple lexical entries or different kind of multi-words (terms, locutions, named entities). These selected lexical units are translated thanks to Wikipedia bilingual dictionary and we get one or several translated candidates for each lexical unit of the query. However, some lexical units don't get any translation at all. For a given query, we keep solutions of segmentation that give the maximum number of translated units and the longest units: we want to give priority to multi-words translation. The second phase is the disambiguation. Since there are often several alternatives for each lexical unit, many combinations can be candidates to the final translation. We choose the best combination according to a criterion of thematic homogeneity (Gledson and Keane, 2008). We use our English semantic resource (shortest category paths) to represent the semantic field of each selected lexical unit. Like for the query expansion task, we transform all the paths linked to a lexical unit to a "bag of categories". Finally, each lexical unit is represented by a

[6] According to the criteria described in section 3.2.

[7] frwiki-latest-langlinks.sql
[8] frwiki-latest-page.sql

flat vector of categories. The proximity between two category vectors is given by a cosine measure. This calculus is done between two adjacent units. In the general case (more than two units in a query), we choose the solution that maximizes the total sum of adjacent units. Figure 2 illustrates the overall process.

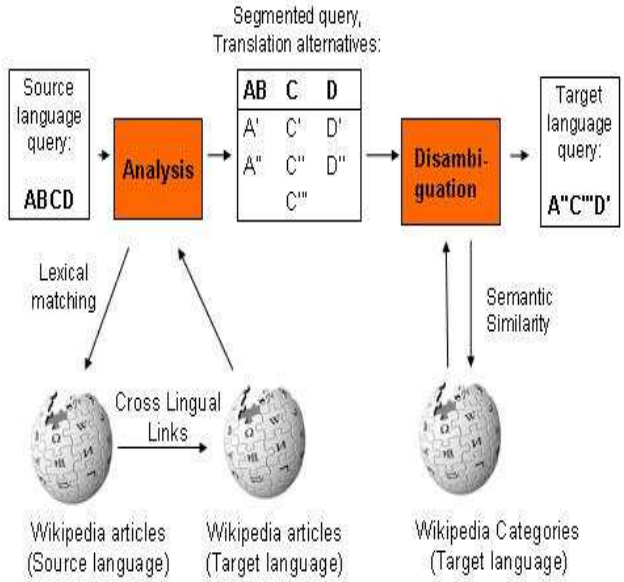

Figure 2: Wikipedia-based query translation. A query consisting of 4 words A, B, C and D is analyzed into 3 lexical units AB, C and D that have several candidate translations. After disambiguation, the A"C'"D' combination is deemed the most consistent.

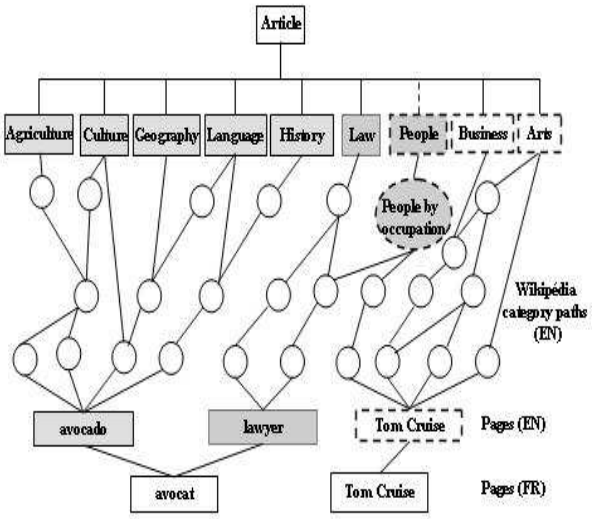

Figure 3: Disambiguation of the query "avocat Tom Cruise"

Figure 3 shows the proximity between selected lexical units of the query "*avocat Tom Cruise*" and their associated shortest path reaching "*Article*" (only pages and terminal categories under "*Article*" are shown). We can see the overlap of "*Tom Cruise*" categories with "*Lawyer*" categories: "*People by occupation*" and "*People*". There is no category overlap with "*Avocado*" categories and "*Tom Cruise*" categories.

4.3 Results

We measured the accuracy of the translation of the prototype on a corpus of 750 queries issued from a monolingual multimedia search engine over three days during November 2009. Many of these 750 queries were typed in on several occasions. So the total number of queries in the corpus is about 7000.

Our goal was to get an idea of the performance of this system compared to standard solutions We compared the translations of these queries by our prototype with the translations of three well known MT services of the market, available online, freely: the online Systran solution[9], the ProMT online application[10], and the Google CLIR service[11]. We evaluate the Error Rate (ER) of each translator on the corpus. Our manual evaluation method was the following: each translation was given an accuracy score of success (0 for a wrongly translated query or not translated at, 0.5 for a partially correct translation and 1 for a good translation). The mean score M is computed over all these scores and weighted by the query frequency. The ER is defined by the formula: ER=1-M. Table 1 gives the following results:

M can be computed based on the 750 queries or based on each occurrence of each query (over the 7000 occurrences). We call it a weighted mean and the resulting ER is a called the weighted ER (ERw). Our prototype has no spelling mistake processing module and no grammatical processing at all either. Therefore, in order to compare its score with the three other state-of-the-art translators, we also measured the ER over the subset of queries that have no spelling mistake and no grammatical feature. For example the query "dog of Obama" would be grammatical because of the "of" genitive marker, as well as plurals. Each MT service or prototype was therefore given 6 different scores: ER over all the queries, ER over all the queries that have no spelling mistake or grammatical feature (ER-sg) and ER over the queries that do have spelling mistakes or grammatical features (ER|sg), these three rates weighted (ERw) or "flat". The results are presented in Table 1 (a lower ER means a more accurate translation):

	"Wiki"	**Systran**	**ProMT**	**Google**
ERw	**0.131**	**0.132**	**0.170**	**0.077**
ER	**0.331**	0.245	0.298	0.177
ERw- sg	**0.100**	**0.118**	**0.156**	**0.064**
ER- sg	**0.175**	0.155	0.225	0.111
ERw |sg	0.713	0.373	0.410	0.286
ER | sg	0.711	0.461	0.477	0.340

Table 1: ER Comparison of various MT solutions.

Several results can be highlighted. On the subset of queries that have no spelling mistake or grammatical feature, our ER is equal or slightly lower than the ER of

[9] http://www.systran.fr/
[10] http://tr.voila.fr/
[11] http://www.google.fr/language_tools?hl=fr

other MT solutions, except Google. Since our system has no spelling or grammatical features, results on the spelling and grammatical queries (ER-sg) show that our prototype is very sensitive to spelling mistakes and grammatical features, its ER-sg query is higher than the others. This result shows that other system probably use spelling or grammatical features.

The fact that our accuracy is consistently much better with the weighted mean accuracy measure means that the most frequent queries are easier for our prototype to translate.

The weighted ER (ERw) keeps all the queries and measures the real performance of our system from the user point of view. Our result (13.1%) is comparable and even better than those of Systran or ProMT standard on line translation solutions, but worse than the specialised Google CLIR solution (nearly half errors). We have shown that our system is penalized in different ways (misspelling, grammatical parser and the lack of bilingual dictionary is far from being exhaustive, especially for standard lexical entries (verbs, common nouns). We think that the quite good performance of our system is partially du to the named entities frequency in our corpus. Nearly 60% of the queries contain a named entity and the Wikipedia bilingual dictionary contains many translated named entities.

We have designed a quite simple query translation system which only relies on Wikipedia data. These data are Wikipedia bilingual dictionary and a "bag of categories" for disambiguation purpose. This system is operational (not yet part of our search engine) and gives performance close to on line standard systems, the more adapted one being Google CLIR service. We have shown some of its weaknesses that we will soon improve.

5. Conclusion

We have filtered the Wikipedia categories lattice by the mean of a shortest path strategy. The result is a sub-lattice which extends each page with several parent category paths. These extensions generalize the semantic of category pages along hyperonymic and thematic axis. These linguistic relations are not explicitly labeled but the generated representation space is useful enough to perform semantic query expansion and query translation disambiguation for queries related to a multimedia search engine. We will soon test its validity on a monolingual disambiguation task. Results seem to confirm our simple winning strategy which supposes that shortest categories paths are the most relevant.

In a future work, we will try to formalize this hypothesis within a theorical framework, the Minimum Description Length theory, which seems to be a logical way to follow. On a second hand we will further compare our resource to existing semantic resources, especially on a linguistic point of view. Automatic labeling of linguistic relations within the extracted sub-lattice is also in the scope of our next work.

6. References

Gaillard, B., Bouraoui, J.L., Guimier de Neef, E., Boualem, M. (2010). Expansion de requêtes pour la recherche d'information multilingue. *In Proceedings of the Conférence en Recherche d'Information et Applications (CORIA 2010)*.

Girault, T., (2008). Exploitation de treillis de Galois en désambiguïsation non supervisée d'entités nommées. *In Proceedings of 15ème conférence sur le Traitement Automatique des Langues Naturelles, TALN'08*,260--269

Girault, T., (2008). Concept Lattice Mining for Unsupervised Named Entity Disambiguation. Concept *In Proceedings of Lattices and their Applications, CLA'08*, 32--43

Gledson, A., Keane, .J. (2008). Measuring Topic Homogeneity and its Application to Dictionary-Based Word-Sense Disambiguation. *In Proceedings of the 22nd International Conference on Computational Linguistics, Coling 2008, Manchester, UK,* pp 273--280.

Guegan, M. (2006). Catégorisation par les contributeurs des articles de l'encyclopédie Wikipedia.fr. *Mémoire de master de recherche informatique université paris XI, LIMSI CNRS*

Medelyan, O., Legg, C,Milne, D, Witten I.H.,(2008). Mining Meaning from Wikipedia. *Working Paper September 2008*.

Mihalcea, R. (2007). Using Wikipedia for Automatic word Sense Disambiguation. In *Proceedings of the NAACL 2007*, pp 196-203

Nastase, V. Strube, M. (2008). Decoding Wikipedia catégories for knowledge acquisition, *In Proceedings of the ,23rd national conference on Artificial intelligence (AAAI 2008)*, 1219-1224

Peng Y., Mao M. (2008), Blind Relevance Feedback with Wikipedia: Enterprise Track, *Proceedings of The Seventeenth Text REtrieval Conference (TREC 2008)*, 18-21

Ponzetto, S.P., Strube, M. (2007). Deriving a large scale taxonomy from Wikipedia. AAAI'07. *In Proceedings of 22nd national conference on Artificial intelligence*, 1440-1445.

Schönhofen, P., Benczur, A., Biro, I. and Csalogany, K. (2008). Cross-Language Retrieval with Wikipedia. *Advances in Multilingual and Multimodal Information Retrieval, Revised selected paper of CLEF 2007, Springer*, 72-79

Strube M., Ponzetto S. P. (2006). WikiRelate!: Computing Semantic Relatedness Using Wikipedia. *In proceedings of AAAI 2006*, 1419-1424

Suchanek, F.M., Kasneci, G., Weikum, G., (2008). Yago: A large Ontology from Wikipedia and WordNet. *Journal of Web semantics,Elsevier,*203-217

Tien-Chien L., Shih-Hung W.(2008), Query Expansion via Link Analysis of Wikipedia for CLIR, *Proceedings of NTCIR-7 Workshop Meeting*, 125-131

Zesch., T., Gurevych, I.,(2007). Analysis of the Wikipedia Category Graph for NLP Applications. *In proceedings of*

the Workshop TextGraphs-2:Graph-Based Algorithms for Natural Language Processing, 1-8

Zesch., T., Gurevych, I. and Mühlhäuser, M. (2007). Analysing and Accessing Wikipedia as a Lexical Semantic Resource. *In Data Structures for Linguistic Resources and Applications*, 197-205

The Automatic Identification of Conceptual Metaphors in Hungarian Texts: A Corpus-Based Analysis

Anna Babarczy[1], Ildikó Bencze M.[1], István Fekete[1], Eszter Simon[1,2]

[1]Budapest University of Technology and Economics
Department of Cognitive Science
H-1111 Budapest, Stoczek u. 2.
E-mail: {babarczy, ibencze, ifekete, esimon}@cogsci.bme.hu
[2]Research Institute for Linguistics, Hungarian Academy of Sciences
H-1068 Budapest, Benczúr u. 33.
E-mail: eszter@nytud.hu

Abstract

The present study is a corpus-based analysis of literal versus metaphorical language use. Previous corpus linguistic works have focused on the linguistic characteristics of the metaphorical expressions. The main question of the present paper is whether the automatic identification of certain conceptual metaphors could be successful taking the embodiment hypothesis as a starting point. 12 widespread conceptual metaphors were selected from Lakoff & Johnson (1980) and the metaphor index in Kövecses (2002), where consistent mapping was observed between a concrete (source) domain and an abstract (target) domain. According to our hypothesis, a metaphoric sentence should include both source-domain and target-domain expressions. This assumption was tested relying on three different methods of selecting target-domain and source-domain expressions: a psycholinguistic word association method, a dictionary method and a corpus-based method. The results show that for the automatic identification of metaphorical expressions, the corpus-based method is the most effective strategy, which suggests that the concept of source and target domains is best characterized by statistical patterns rather than by psycholinguistic factors.
Keywords: embodiment hypothesis, conceptual metaphors, association, corpus-based, automatic identification

1. The Theory of Metaphor

1.1 The Cognitive Theory of Metaphor

In everyday language use the term *metaphor* is held to be a figure of speech which refers to an analogy between two entities or concepts (e.g., *Achilles was a lion*). In cognitive linguistics, in contrast, metaphor is first of all a conceptual process, thus metaphorical relations are taken to be conceptual mappings, which characterize not only our language use but also our everyday life, thought and behavior (Lakoff & Johnson, 1980). According to the cognitive linguistic view, conceptual metaphors refer to the understanding of an abstract concept, also called the *target domain*, in terms of a concrete concept of which we can have direct sensory experience, namely the *source domain*. This underlying association between the two domains is held to be systematic in both language and thought.

The hypothesis that the representation of abstract concepts in the mind/brain is grounded in the representation of concrete knowledge, which in turn is grounded in our bodily experience of the world, is the main statement of the **embodiment theory** in cognitive linguistics (Gibbs, 2006; Kövecses, 2002; Lakoff & Johnson, 1980, 1999). For example, people universally think and talk about the abstract concept of "time" with the help of "space", the terms of which are acquired through our interaction with the environment (*before, after, under, in* etc.). Consequently, we can argue that the concept of "time" is structured by the concept of "space" which means that there is a TIME IS SPACE conceptual metaphor in our mind.

This hypothesis is supported by psycholinguistic experiments: it has been shown, for instance, that sensory-motor experiences influence the interpretation of metaphorical expressions on "time" (Boroditsky & Ramscar, 2002) which means that during the understanding of metaphors people do physical *motion simulation*, i.e. they imagine the actions or events described by metaphorical expressions (Gibbs & Matlock, 2008). However, other experiments did not find evidence for the necessity of conceptual metaphoric mappings in comprehension of metaphorical expressions (Keysar et al., 2000; Szamarasz, 2006). The problem whether in natural language use abstract concepts are independent of concrete concepts still remains an open question.

1.2 The Statistical Learning Theory

Another approach referring to the nature of abstract knowledge is the **statistical learning theory,** which argues that people acquire and structure their abstract concepts with the help of the statistical properties of language (Burgess & Lund, 1997; Landauer & Dumais, 1997). This means that novel linguistic symbols are directly abstracted from known symbols without the interference of metaphorical processes or embodied schemes.

The two theoretical approaches do not necessarily exclude one another since it is conceivable that our abstract knowledge exploits both sources mentioned above. According to this integrative point of view (Andrews et al., 2005, 2007), both the attributive and distributive properties of words play an important role in symbol grounding. Attributive properties are non-linguistic physical attributes associated with a word, while

distributive factors refer to common occurrences of a word with other linguistic elements.

Based on our discussion so far, the present paper investigates whether the automatic extraction of conceptual metaphors in large corpora could be successful taking the embodiment hypothesis as starting point, and along with this, whether which strategy is the most effective: the psycholinguistic word association method or the corpus-linguistic method based on statistical patterns.

2. Metaphor and Corpus Linguistics

2.1 Corpus-Based Research on Metaphor

Corpus-based studies of metaphorical language use have already pointed out the inadequacy of the cognitive theory and also the defects of psycholinguistic experiences. These critics claim that the theoretical and experimental research neglect the linguistic attributes of metaphorical expressions, and they do not use natural data but fictitious examples, which might be misleading in some cases. For example, Deignan (2008) demonstrates that according to corpus-linguistic results the conceptual metaphor AN ANGRY GROUP OF PEOPLE IS A WILDFIRE is more likely to occur than the metaphor ANGER IS THE PRESSURE OF HEATED FLUID IN A CONTAINER, even though it is the latter that is ubiquitously listed in works in cognitive theory. Observed metaphorical patterns (Stefanowitsch, 2006) and collocations (Deignan, 2005, 2008) also have characteristic grammatical features. Similarly, Deignan (2005) demonstrates that in metaphoric usage the words have less grammatical liberty compared to their literal occurrences. For example, the words belong to the source domain in the metaphorical mapping tend to denote actions and properties, and thus they occur mainly as verbs and adjectives. These results show that the logical relations between concrete entities are not simply mirrored in abstract language use but undergo some kind of change. This fact supports the so-called *blending theory* (Fauconnier & Turner, 2002), which contends that during metaphoric language use people create a mixed or blended domain that has a proper structure and relations, and thus proper linguistic features.

Taking all the evidence into account, it is clear that the conceptual theory of metaphor alone is not able to explain all the phenomena found in texts.

2.2 Methodological Problems in Automatic Conceptual Metaphor Identification

The default method of metaphor annotation is *manual processing*: based on their linguistic intuitions, researchers mark expressions that they perceive as metaphorical in a given corpus. Since this method is very labor-intensive and time-consuming, it is worth experimenting with at least partly automated techniques, such as *searching a corpus for expressions belonging to the source domain* (e.g., Deignan, 2008) or to *the target domain* (Stefanowitsch, 2006) and manually checking the extracted sentences for metaphoricity. Finally, it is also possible to search the corpus for sentences containing characteristic words from *both the source and the target domains* of a given conceptual metaphor (e.g., Martin, 2006). The disadvantage of this method is that in this way we can test only predetermined metaphorical mappings, and, in contrast to the technique used by Stefanowitsch (2006), the recovery of novel metaphors is precluded. However, it has the advantage of a higher level of automation in the annotation process allowing the processing of larger corpora. It is this latter strategy that our study attempts to enhance.

The first step of any of the above three (semi-) automated methods is that expressions that are likely to characterize either the source domain or the target domain of a given metaphor type need to be collected. However, the identification of the linguistic cues that may characterize a particular domain is not a straightforward question. A problem facing automatic metaphor annotation is that, in general, the domains of conceptual mappings discussed in the cognitive literature are associated with concepts rather than specific linguistic forms. Our paper undertakes to address this issue by testing three different methods of compiling word lists characterizing the source versus the target domains of a set of conceptual metaphors. The first two methods rely on experimental psycholinguistic evidence and on lexicographic data, while the third approach is based on the manual analysis of a reference corpus. In addition to the practical import of the results for corpus analysis, the experiments also shed light on the language theoretical issue discussed in Section 1. If either of the first two methods proves to be more successful, we have some support for the embodiment hypothesis. If, however, the third method leads to the best results, the statistical approach to metaphor proves to be more plausible.

3. The Study: Automatic Identification of Metaphors

The main question addressed by the present study is, therefore, whether the automatic identification of certain conceptual metaphors is feasible taking the concept of source-to-target domain mapping as a starting point.

The experiment involved the following phases:

- A set of conceptual metaphors were selected from the cognitive linguistic literature.
- A corpus was compiled using a variety of text types.
- Word lists characterizing the source and the target domains of the selected conceptual metaphors were compiled using three different methods. This resulted in three separate sets of source-target word lists.
- Sentences containing at least one source-domain word and at least one corresponding target-domain word were automatically extracted from the corpus. The three sets of word lists were used in separate runs.
- The results were manually checked for precision and recall.

3.1 Resources and Methods

3.1.1. The Conceptual Metaphors

12 widespread conceptual metaphors were selected from Lakoff & Johnson (1980) and the metaphor index in Kövecses (2002). The criteria for the selection process were the following:

- The metaphor had to be general enough to be found in many types of texts,
- The domains had to be suitable for providing associations in a psycholinguistic experiment, and
- There had to be a mapping from a concrete source domain to an abstract target domain.

Based on the above, the following 12 conceptual metaphors were chosen:

1. ANGER IS HEAT
2. CHANGE IS MOTION
3. CONFLICT IS FIRE
4. CONTROL IS UP
5. CREATION IS BUILDING
6. MORE IS UP (LESS IS DOWN)
7. POLITICS IS WAR
8. PROGRESS IS MOTION FORWARD
9. RESOURCES ARE FOOD
10. THE MIND IS A MACHINE
11. THEORIES ARE BUILDINGS
12. TIME IS MONEY

3.1.2. The Corpus

The corpus was compiled observing two criteria: a variety of genres should be represented; and the texts should be accessible for research purposes in four different languages. The genres include modern fiction from digital libraries, popular science articles from the National Geographic magazine and movie subtitles, the latter of which was included as a representation of quasi-spoken language. The criterion of multilingual availability was needed in view of future plans of creating a multilingual parallel corpus (Hungarian, English, Spanish and Italian) with metaphor annotation. As the analysis has only been completed for the Hungarian texts, the results described in this paper apply to the Hungarian corpus. The sizes of the Hungarian texts from the different genres are shown in Table 1.

Text types	**Number of text words**
National Geographic	68,997
Subtitles	32,148
Fiction	208,384
Total	**309,529**

Table 1: The content of the corpus.

The texts were converted to plain text format with UTF-8 character encoding. The part-of-speech tagger *Hunpos* (Halácsy et al., 2007) was used to tag the Hungarian texts. Hunpos was chosen because it is a Hidden Markov Model-based open source part-of-speech tagger, which can tag any language once it has been trained on a pre-tagged corpus. As the next step, the tagged corpus was converted to XML format, which was our working format for metaphor identification.

3.1.3. The Baseline Corpus

In order to obtain an estimate of the performance expected from an automatic metaphor annotation method a baseline corpus was constructed on which human inter-annotator agreement was measured.

The baseline corpus was created by extracting 10% (approximately 30,000 words) of the entire corpus in which each genre was represented in the same proportion as in the main corpus. The baseline corpus was independently annotated for metaphors by two human annotators.

The manual annotation followed a pre-defined procedure. The procedure was based on the criteria defined by Pragglejaz (2007). For example, classical idioms, i.e., fixed collocations which are not decomposable (e.g., *pop the question*), "dead metaphors" or those which are metaphorical only in etymological sense (e.g., the word *depression*) were not classed as metaphorical. A rule was further defined for each type of conceptual metaphor. For example, in the case of the MORE IS UP conceptual metaphor we applied the following rules: "Every expression with a 'quantity' meaning which can be visualized as moving along a vertical scale, e.g., *price, lease, temperature,* should be annotated as a potential target domain expression. Every sentence which contains the word *csúcs* ('top') e.g., *csúcsteljesítmény* ('top performance'), *csúcstechnológia* ('peak technology') should be annotated as metaphorical."

At the first attempt, inter-annotator agreement was only 17%. After refining the annotation instructions, we made a second attempt, which resulted in an agreement level of 48%, which is still a strikingly low value. These results indicate that the definition of "metaphoricity" is problematic in itself.

Some typical sources of disagreement between the annotators are the following:

- In the absence of a statistical measure of semantic distance, it was difficult to draw the line between words directly referring to a concept belonging to the source domain and those indirectly referring to it. For example, in the case of the conceptual metaphors ANGER IS HEAT or CONFLICT IS FIRE, the source domain should be an expression referring to a sort of "heated thing". However, in some cases, one or the other annotator included words indirectly suggesting the presence of heat, such as *kiolt* ('extinguish'), *kihűl* ('get cold') etc. Another case in point is the phrase *a memória élesítése* ('the sharpening of one's memory'), which may or may not be an instance of the conceptual metaphor THE MIND IS A MACHINE, depending on whether the annotator is prepared to accept the indirect association between machines and acts of sharpening.
- A second source of discrepancies was the fuzzy nature of the boundary between ambiguous words having an

established abstract sense and metaphorical uses of unambiguous words. For example, the expression *eljutottam a mai napig* ('I've gotten to this day') may or may not represent a CHANGE IS MOTION metaphor depending on whether the Hungarian verb *jut* (literally: get somewhere, reach a place by moving the entire body) is taken only to denote physical movement or to be ambiguous. The verb *alapul* ('be founded on something'), which is derived from the noun *alap* ('foundation') is similarly problematic since, although *az elmélet alapjai* ('foundations of the theory') is a good example for THEORIES ARE BUILDINGS, the verb derived from the concrete noun can only have an abstract sense. The question is, therefore, how far we should go in diachronic or morphological analysis when making a decision of metaphoricity.

- The level of inter-annotator agreement was further lowered by discrepancies in the classification of metaphorical expressions. Consider the following example from the novel *The Master and Margarita*: *az öreg előbb **megdöntötte** mind az öt **bizonyítékot**, és aztán, mintegy magamagából csúfot űzve, ő maga **felállított** egy hatodikat.* ('the old man first **demolished** all five **arguments** and then, as if mocking himself, **constructed** a sixth of his own'). This phrase were classified by one of the annotators as a THEORIES ARE BUILDINGS metaphor, while the other considered it to pertain to a CREATION IS BUILDING type. Similarly, it is difficult to make an informed decision on whether the following example contains a CHANGE IS MOTION or a PROGRESS IS MOTION FORWARD metaphor, neither of which appear to be an intuitively correct choice: *a járvány végigsöpört szülővárosukon* ('the epidemic swept through their hometown').

3.1.4. The Compilation of the Word Lists

For the automatic identification of metaphors, we searched the corpus for sentences containing one or more words characterizing the source domain and one or more words representing the target domain of a given conceptual metaphor. Three different methods of compiling the word lists were tested: a) word association experiment, b) dictionary of synonyms, and c) reference corpus.

The first method is based on the assumption that the expressions people associate with a key word for the source domain and a key word for the target domain can provide a lexical profile for a given metaphor type. The word associations were collected in an online experiment. 138 students from the Budapest University of Technology and Economics participated in the experiment. One key word for each source and target domain (e.g., *anger, building, change, up, war*) appeared on the screen one at a time in randomized order and the participants had one minute to type words they associated with the key word. When the minute was up, the keyword disappeared and participants were instructed to click a button when they were ready for the next key word.

The lists obtained in the association experiment were normalized: multiword expressions, proper names and antonyms were filtered out, abbreviations were completed, and finally, the words were stemmed by the Hunmorph open source morphological analyzer (Trón et al., 2005).

For each of the 12 conceptual metaphors, the resulting two word association lists (one containing associations provided for the source domain, and another providing associations for the target domain) were taken to constitute the metaphor's lexical profile.

For the second method, the word lists obtained from the association experiment were expanded with the synonyms listed for the association words in the *Magyar szókincstár* [Hungarian Word Thesaurus] (Kiss, 2007). Dialectal, slang and obsolete expressions were omitted. Compared to the association list, the size of the word lists substantially increased (see Table 2). For the third -- corpus-based -- method, the word lists for each source and target domain were extracted from the manually annotated baseline corpus. Due to the low level of inter-annotator agreement obtained for the baseline corpus, the union of sentences annotated as metaphorical by the two annotators were used for compiling the corpus-based lists of source and target domain words.

Method / Words	Psycho-linguistic	Synonyms	Corpus-based
Source domain	1239	6348	126
Target domain	674	5094	120

Table 2. Number of words in source- and target-domain lists compiled by the three methods.

3.1.5. The Annotation Process and its Verification

Based on the three sets of word lists, the XML test corpus was automatically annotated producing three files in which the sentences were marked with tags showing the type of conceptual metaphor the system identified. Each of the three annotation versions were then verified manually using the graphical interface of the GATE application (Cunningham et al., 2002). Because of time constraints, the manual verification was completed for 10% of the test corpus, where the different genres were represented in the same proportion as in the entire corpus. In this sub-corpus, a total of 155 sentences were identified as metaphorical by two human annotators.

3.2 Results

The results of the three methods were quantified by the *precision* and *recall* measures (Table 3). Precision shows the proportion of the sentences correctly tagged as metaphorical by the automatic system, while the recall measure shows the percentage of metaphorical sentences successfully identified by the system. The F-measure is

the weighted harmonic mean of these values, i.e. the final indicator of the system's performance.

Method	Recall	Precision	F-measure
Association	3.8%	7.5%	5.6%
Dictionary	18.1%	4.5%	11.3%
Corpus	31.3%	55.4%	43.3%

Table 3: Results of the three methods.

The results reveal that the association method covered substantially fewer metaphorical sentences containing both a source and a target expression than the other two methods. This psycholinguistic method also performed very poorly in terms of precision. When the association word lists were expanded with synonyms, recall somewhat improved but only at the cost of a decline in precision. The corpus-based method was very clearly the most successful of the three strategies. Taking all our results into account, we must contend that the hypothesis that the co-occurrence of psycholinguistically typical source domain and target domain words in a sentence is a good predictor of metaphoricity receives no empirical support. Exploiting the statistical properties of texts leads to considerably better but still not satisfying results.

3.3 Problem Cases

It is clear from the above discussion that deciding whether a sentence is metaphorical or not is far from being a straightforward task. The general experience of our experiments is that if certain elements are difficult for a human language user to find in a text, then the automatic identification of these words also brings poor results. One problem is that in several cases we must look beyond a single sentence. The manual annotation identified several sentences that were metaphorical but did not contain words from both the source and the target domains, i.e. they were problematic with regard to recall. There were sentences in which a word denoting a concrete action in its literal interpretation (source domain) referred to a metaphorical event, which could only be deduced from the extra-sentential context.

In other cases, the metaphoricity of the sentence was signaled by a single word which incorporated both the source and the target meaning.

Precision values were lowered by the frequent occurrence of sentences which contained both a source and a target expression but were not metaphorical. A typical example is given below: *Mérnökök és **vezetők** tanakodnak kisebb csoportokban a 23 emelet **magas** fúrótorony tövében.* ('Small groups of engineers and **managers** are discussing their options at the base of the 23-storey **tall** oil-rig.')

The word *manager* is a target-domain expression and the adjective *tall* is a source-domain expression for the metaphor CONTROL IS UP but the two words are conceptually unrelated in this particular sentence.

4. Conclusions

The present paper investigated the automatic identification of conceptual metaphors using corpus-linguistic analyses, and found that the concept of source and target domains is best characterized by statistical patterns rather than by psycholinguistic factors. Since the main objective of our study was to find the most effective way of automatically identifying conceptual metaphors in natural texts, we did not carry out a detailed grammatical analysis of the examples or explore the possible connection between the type of texts and the type of metaphors occurring in them. However, it seems that our research supports previous results of corpus-linguistic analyses, in particular those regarding collocations and the linguistic form of metaphorical expressions. This is also confirmed by the fact that, while the lists compiled on the basis of the association experiment had a very weak predictive force, the targeted selection of the words characteristic to conceptual domains brought the best result, which means that not every association suggests metaphoricity but only the common co-occurrences of certain words and expressions. For example, in Hungarian the co-occurrence of the verb *pazarol* ('waste') and the noun *idő* ('time'), or the verb *gerjeszt* ('induce') and the noun *harag* ('anger') within a single sentence almost always signals a metaphor.

Our analyses also found several examples highlighting the importance of grammatical form: for example, in the case of the conceptual metaphor RESOURCES ARE FOOD, according to the reference corpus method the source domain is represented mainly by verbs (*fogyaszt* 'consume', *felfal* 'devour', *táplál* 'feed'), while the majority of words collected in association experiment are nouns (*edény* 'dish', *fagylalt* 'ice cream', *reggeli* 'breakfast' etc.). This observation supports the results obtained by Deignan (2005) showing that for the majority of metaphorical expressions, words referring to the source domain are verbs or adjectives. The author argues that this is because in metaphorical language use people try to describe abstract entities, thus they take words denoting behaviors, features or actions from the concrete source domains. Of course, the confirmation of these hypotheses requires a more comprehensive analysis of the metaphors found so far. Our plans for the future involve the expansion of the reference corpus and the extraction of a larger word list for source and target domains. At the same time, we intend to analyze the English, Spanish and Italian versions of the texts, and to compare the results with the Hungarian data, since cross-linguistic analyses might reveal important factors in the conceptual nature of metaphorical expressions.

5. References

Andrews, M., Vigliocco, G., Vinson, D. (2005). The role of attributional and distributional information in semantic representation. In B. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the Twenty Seventh Annual Conference of the Cognitive Science Society*.

Andrews, M., Vinson, D., Vigliocco, G. (2007). Evaluating the Contribution of Intra-Linguistic and Extra-Linguistic Data to the Structure of Human Semantic Representations. In *Proceedings of the Cognitive Science Society*.

Boroditsky, L., Ramscar, M. (2002). The roles of body and mind in abstract thought. *Psychological Science*, 13, pp. 185–188.

Burgess, C., Lund, K. (1997). Representing abstract words and emotional connotation in high-dimensional memory space. In *Proceedings of the Cognitive Science Society*, Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 61–66.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V. (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia.

Deignan, A. (2005). *Metaphor and corpus linguistics*, Amsterdam/Philadelphia: John Benjamins.

Deignan, A. (2008). Corpus linguistics and metaphor. In R.W. Gibbs Jr. (Ed.), *The Cambridge Handbook of Metaphor and Thought*, Cambridge: Cambridge University Press, pp. 280–294.

Fauconnier, G., Turner, M. (2002). *The way we think: conceptual blending and the mind's hidden complexities*. New York: Basicbooks.

Gibbs, R.W. (2006). *Embodiment and cognitive science*, New York: Cambridge University Press.

Gibbs Jr., R.W., Matlock, T. (2008). Metaphor, imagination and simulation. Psycholinguistic evidence. In R.W. Gibbs Jr., (Ed.), *The Cambridge Handbook of Metaphor and Thought*, Cambridge: Cambridge University Press, pp. 161–176.

Keysar, B., Shen, Y., Glucksberg, S., Horton, W.S. (2000). Conventional language: How metaphorical is it? *Journal of Memory and Language*, 43, pp. 576–593.

Kiss, G. (2007). Magyar szókincstár [Hungarian Word Thesaurus], Budapest: Tinta.

Kövecses, Z. (2002). *Metaphor. A Practical Introduction,* Oxford: University Press.

Lakoff, G., Johnson, M. (1980). *Metaphors we live by*, Chicago: University of Chicago Press.

Lakoff, G., Johnson, M. (1999). *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*, New York, NY: Basic Books.

Landauer, T.K., Dumais, S.T. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2), pp. 211–240.

Martin, J.H. (2006). A corpus-based analysis of context effects on metaphor comprehension. In A. Stefanowitsch,& S.Th. Gries (Eds.), *Corpus-based approaches to metaphor and metonymy*, Berlin/New York: Mouton de Gruyter, pp. 214–236.

Halácsy, P., Kornai, A., Oravecz, Cs. (2007). HunPos - an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, Association for Computational Linguistics, Prague, Czech Republic, pp. 209–212.

Pragglejaz Group. (2007). MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1), pp. 1–39.

Stefanowitsch, A. (2006). Words and their metaphors: a corpus-based approach. In A. Stefanowitsch & S.Th. Gries (Eds.), *Corpus-based approaches to metaphor and metonymy,* Berlin/New York: Mouton de Gruyter, pp. 63–105.

Szamarasz, V.Z. (2006). Az idő téri metaforái: a metaforák szerepe a feldolgozásban. *Világosság*, 47(8-9-10), pp. 99–109.

Trón, V., Németh, L., Halácsy, P., Kornai, A., Gyepesi, Gy., Varga, D. (2005). Hunmorph: open source word analysis. In *Proceedings of the ACL 2005 Workshop on Software*, pp. 77–85.

Towards a Name Entity Aligned Bilingual Corpus

Xiaoyi Ma
Linguistic Data Consortium
3600 Market St. Suite 810
Philadelphia, PA 19104
E-mail: xma@ldc.upenn.edu

Abstract

This paper describes a co-training framework in which, through named entity aligned bilingual text, named entity taggers can complement and improve each other via an iterative process. This co-training approach allows us to 1) apply our method to not only parallel but also comparable text, greatly extending the applicability of the approach; and to 2) adapt named entity taggers to new domains; 3) create a named entity aligned bilingual corpus. Experiment results on Chinese-English data are shown and discussed.

1. Introduction

Named entity aligned bilingual corpora are valuable resources for many NLP applications, including machine translation, cross-lingual information retrieval. Manually annotating such corpora is extremely expensive, time consuming, and it cannot be scaled up easily, which makes automatic creation of these corpora a very attractive approach, given the amount of bilingual text that becomes available everyday.

Automatic bilingual named entity alignment usually involves two steps: 1) identification of names in both halves of the bitext; 2) alignment of names across two languages.

Automatic bilingual named entity alignment faces a couple of difficulties. First and foremost, current state-of-art named entity taggers don't adapt well to new domain and time epochs. Rule-based (Grishman, 1995) and statistical named entity tagging methods, such as hidden markov models (Bikel et al., 1999), maximum entropy models (Borthwick, 1999), and conditional random fields (Li and McCallum, 2003), performs well in the targeting domain, but there performance decreases significantly on data from other domains or time epochs.

Secondly, alignment of names across languages can be tricky due to a number of reasons: 1) name translation and transliteration variations; 2) named entities can be ambiguous, that is, the same "name" can refer to different entities.

This paper is part of our ongoing research on named entity alignment on unlabeled bilingual text, which has the following major goals:

a) improve current state-of-the-art taggers;
b) adapt existing taggers to new domains;
c) automatic alignment of named entities in bilingual texts;

This paper focuses on improving named entity taggers using unlabeled bilingual text and to adapt these taggers to new domains within a co-training framework. Work on aligning the named entities is yet to be completed. However, preliminary experiment result demonstrates that the taggers have good coverage and accuracy on the bitext we're about to conduct entity alignment on, which lays a solid ground for future work on named entity alignment.

This paper is laid out as follows. Section 2 provides the background for this paper. Section 3 describes the entity alignment-based co-training algorithm for enhancing NE taggers, as well as the general approach of entity alignment in parallel and comparable text. Section 4 describes the experiments done on English-Chinese parallel text and comparable text. Section 5 shows the experiment results. Section 6 concludes this paper.

2. Previous Works

Previous works on inducing or enhancing text analysis tools using bilingual text include (Yarowsky et al., 2001) and (Hwa et al., 2005).

(Yarowsky et al., 2001) describes a set of algorithms for automatically inducing text analysis tools – POS taggers, base noun-phrase bracketers, named entity taggers, and morphological analyzers – for an arbitrary foreign language from English, using aligned parallel text corpora. Parallel text corpora were first word/character aligned using the EGYPT system (Al-Onaizan et al., 1999). The English side of the corpus is tagged or bracketed using the state-of-the-art taggers or bracketers, and the English tags/brackets are then projected to the foreign language. Since the direct annotation projection is noisy, the paper presents training procedures and algorithms to bootstrap taggers from noisy and incomplete initial projection.

To induce named entity taggers from aligned parallel text corpora, (Yarowsky et al., 2001) did the initial classification on a per-word basis, using an aggressively smoothed transitive projection model. The co-training-based algorithm given in (Cucerzan and Yarowsky, 1999) was then used to train a named entity tagger from the projected data. To evaluate the performance of the induction algorithm on named entities, (Yarowsky et al., 2001) used the Canadian Hansard corpus with about 2.8M sentence pairs, the English side was first tagged by a tagger trained on MUC-6 training data, then the tags were projected to the French side and the projected data were used to train a French named entity tagger. The named entity

tagger achieved 85% classification accuracy measured in terms of per-word entity-type classification accuracy on 4 entity types: FNAME, LNAME, PLACE, and OTHER. The paper claims the induced French tagger is near perfect since the original English tagger achieved only 86% accuracy.

(Hwa et al., 2005) adopted a similar approach to bootstrap non-English syntactic parsers from English by using a state-of-the-art English parser and parallel text. The English side of the parallel text is first analyzed using the state-of-the-art parser, the parse trees are then converted to dependency structures, which are projected across the word alignment to the non-English side using a direct project algorithm. To address the structural differences between English and non-English languages, (Hwa et al., 2005) apply a small set of manually compiled, language-specific post-projection transformation rules on the projected trees. Finally, (Hwa et al., 2005) uses aggressive filtering strategy to automatically prune out projected trees that are believed to be of poor quality. The resulting trees are then used to train a new dependency parser.

Co-training (Blum and Mitchell, 1998) assumes features can be partitioned into two different sets to represent different views of the same data, and in addition, it assumes each view by itself would be sufficient for learning if there were enough labeled data. Initially two separate classifiers are trained with labeled data. Each classifier was then used to classify the unlabeled data and each classifier's prediction on the unlabeled data is used to augment the training set of the other. Each classifier is retrained with the additional training data provided by the other classifier, and the process repeats.

(Blum and Mitchell, 1998) applied the co-training algorithm to web page classifiers which are trained to identify course web pages from a set of web pages collected from Computer Science department websites at four universities. Three naive Bayes based classifiers were trained on the labeled data, one page based, one hyperlink based, and the third page-hyperlink combined. Experiment results show the co-training algorithm improves all three classifiers significantly, and in the case of combined classifier, the co-training algorithm was able to reduce the error rate by more than 50%.

3. The Co-training Algorithm

Named entity tagging in the context of bilingual text fits the co-training framework nicely. Bilingual texts of the same content (news event, biomedical paper, etc) are naturally two views of the same data. Each view is sufficient for learning of named entity tagging, given enough labeled data.

Figure 1 illustrates the co-training algorithm, which utilizes parallel text to improve NE taggers, using English and Chinese as an example.

In essence, the algorithm iteratively selects new training instances from unlabeled text to augment labeled training data. During the initialization stage, both sides of the parallel text are labeled by the baseline taggers trained on labeled training data, $E_{labeled}$ and $C_{labeled}$ (lines 2 to 5). On each iteration, using labeled English data for supervision, the algorithm selects Chinese data that the current Chinese NE tagger fails to label correctly, and these data (with their labels corrected) are used to augment the training data for Chinese (line 13). The augmented training data is used to train a new and better Chinese named entity tagger (line 14). The new Chinese tagger is then used to re-tag the Chinese text (line 16). Using the newly tagged Chinese text for supervision, English training data is augmented (line 18) and used to train a new English tagger (line 19). And the process repeats for N iterations.

1 Initialization
2 train English NER model $Etagger_{baseline}$ on $E_{labeled}$
3 train Chinese NER model $Ctagger_{baseline}$ on $C_{labeled}$
4 $ETagged_{baseline} \leftarrow$ tag English side of the parallel text using $Etagger_{baseline}$
5 $CTagged_{baseline} \leftarrow$ tag Chinese side of the parallel text using $Ctagger_{baseline}$
6 $ETagged_{latest} = ETagged_{baseline}$
7 $CTagged_{latest} = CTagged_{baseline}$
8
9 **For *i* in *1* to *N***
10 $Ctrain_{add} \leftarrow \phi$
11 $Etrain_{add} \leftarrow \phi$
12
13 $CTrain_{add} \leftarrow augE2C(ETagged_{latest}, CTagged_{baseline})$
14 train Chinese NER model $Ctagger_i$ on combine($C_{labeled}$, $CTrain_{add}$)
15
16 $CTagged_{latest} \leftarrow$ tag Chinese side of the parallel text using $Ctagger_i$
17
18 $ETrain_{add} \leftarrow augC2E(ETagged_{baseline}, CTagged_{latest})$
19 train English NER model $Etagger_i$ on combine($E_{labeled}$, $ETrain_{add}$)
20
21 $ETagged_{latest} \leftarrow$ tag English side of the parallel text using $Etagger_i$
22 done

Figure 1 Co-training algorithm for English and Chinese named entity taggers

Given both sides of the parallel text with automatic labels, functions augE2C and augC2E augment Chinese and English training data respectively by

projecting tags from English to Chinese and Chinese to English.

3.1. Filtering Noises

Noises may be added to the training data and propagates, leading to the deterioration of the NE taggers' performance. The noises come from two sources: 1) incorrectly labeled tokens on both sides of parallel text; 2) the name projection process, which can project correctly labeled tokens incorrectly across languages.

The noise filtering approaches we adopted include local and global validation, and orthography-based filtering. Local and global validation validates strings that are identified as names. Orthography-based filtering makes sure all names in a sentence have been identified.

Local and global validation – For any given name label pair (*n, t*) where *n* is a name and *t* is the type of the name, local and global validation seeks supporting evidence that *t* is the correct label for *n*. If (*n, t*) fails both local and global validation, the word label pair would be deemed unreliable, it wouldn't be used for tag projection, and sentences containing the word would be disqualified as new training examples.

Local validation of (*n, t*) passes if there is at least another instance of string *n* within the same document and if all instances of name *n* bear the same label *t*. Local validation is based on the hypothesis that within a document the name type of the same name should be highly consistent.

If a name fails local validation (either because there aren't other instances in the same document, or the instances of the name aren't labeled consistently), global validation would decide if the name and type are valid. Global validation considers how a name is labeled in the entire corpus. If the label consistency of a name exceeds a preset threshold (85% in all experiments in this paper), the name and type would be considered valid.

If we look at each iteration of the co-training algorithm as a two-step process, in the first step to project names from one language to another, and in the second step to select sentences to augment the training data, then the name validation can be applied to both steps. Before a name is projected to the other language, the name has to be validated either locally or globally so that we're fairly confident with the label. Also, before a sentence is added to the training data, we validate all the names in the sentence. A sentence would be disqualified as new training data if any name in the sentence fails both local validation and global validation.

Orthography-based filtering – local and global validation filters out names that are identified but incorrectly labeled (for example, *George Bush* labeled as an organization). Another type of noise is those names that aren't identified at all (those labeled as *O* in BIO scheme). Sentences containing unidentified names should not be used as new training examples. In languages that exhibit orthographic differences between names and non-names, such as English, exploring the orthographic differences can effectively filter out sentences containing unidentified names. For example, in case of English, person names, location names and organization names are written with the initial letter of each word capitalized, while most non-names are not. So an aggressive and simple heuristic for filtering out sentences with unidentified names is to discard all sentences containing words (except the first word in a sentence) with the initial letter capitalized but not identified as a name.

We compiled a capital_non_name list from ACE 2007 English data. The list consists non-names that are usually written with the first letter capitalized, including job titles, days of a week, months of a year, and names of other types, for example books, movies, drugs.

The orthography-based filter disqualifies English sentences containing word(s) that satisfy all three conditions as follows:

1) the initial letter of the word is capitalized (except when the word is the initial word of the sentence);
2) the word is labeled as a non-name;
3) the word is not on the capital_non_name list;

This procedure inevitably filters out some good sentences as well, which isn't a big concern to us because we have large quantities of unlabeled data.

Orthography-based filtering cannot be applied to languages such as Chinese and Arabic, which don't distinguish names from non-names orthographically. It is still effective if the language pair involves one language that does have orthographical differences between names and non-names.

3.2. Maintaining Data Distribution

A caveat of applying statistical semi-supervised methods, co-training included, is that the new training data extracted from unlabeled data should conform to the underlying data distribution, otherwise the additional training data may skew the statistics and end up hurting the retrained classifier. We choose to use the data distribution in the manually labeled data as the underlying data distribution. The new training data is selected in a way so that it matches the ratio of person names, location names, and organization names in the labeled data.

3.3. Weighting training data

The new training examples extracted from parallel text will undoubtedly contain incorrectly labeled tokens. Naturally the manually labeled data and the extracted sentences should be weighted differently to favor manually labeled data. Between count merging and creating multiple models and calculating weights for each model (model interpolation), (Bacchiani et al., 2006) shows that

count merging is more effective, which is what we employed in our system. We implement count merging by concatenating the training sets, possibly with multiple copies of each to account for weighting.

3.4. Entity Alignment for Chinese-English Bilingual Text

For the co-training algorithm to work, names need to be aligned correctly across bilingual text. If the text is parallel text, alignment can be acquired via automatic word alignment of the parallel text, which is a topic well studied in the context of Machine Translation. The problem with the word alignment approach is two fold. First, it only works on parallel text. Second, it requires large quantities of parallel text to work well.

To achieve named entity alignment, we probably don't need a full-fledged word alignment. There are certain properties of named entity translation that we can take advantage of to achieve high accuracy without aligning every word in the sentence. One observation is that names and non-names are translated differently: names are usually transliterated – with the exception of organization names – while non-names are mostly translated. In addition, large percentage of transliterated names is proper names in the target language that don't overlap with other word categories. These two properties are very effective and quite enough to remove most false positives, as shown in the experiments described in the following sections.

These properties can be explored to align named entities in bilingual document pairs. For parallel text, these pairs are the source text and the translation. In case of comparable text, we use a lexicon based content matching tool to identify document pairs that have similar content.

We employ four approaches to align names, in order of accuracy:

1) Pinyin mapping – a deterministic process to transliterate Chinese into English;
2) Dictionary lookup – looking up possible translation/transliterations from existing bilingual name lists;
3) Transliteration model – use transliteration model trained on transliterated Chinese English name pairs to generate and search for possible transliterations of a name. Models were trained using Moses (Koehn et al., 2007).
4) Google translation – use the Google online translation tool[1] to translate a name.

Some of these methods can be applied to certain entity types only. For example, we don't use transliteration model on organization names, because organization names are usually translated.

[1] http://www.google.com/language_tools

4. Experiments

We first trained baseline Chinese English named entity taggers, then applied the co-training algorithm using Chinese English parallel and comparable text.

4.1. The Data

The baseline taggers were trained on the Chinese and English data from ACE 2005 Multilingual Training Corpus(Doddington et al., 2004). The Chinese training data contains about 308K characters, the English about 190K words. The ACE test data contains about 74K Chinese characters and 58K English words.

The CRF based taggers which identify person, location and organization names uses features such as unigram, bigram, trigram, and pre-defined lexicons.

The parallel text used in this experiment is the FBIS data[2], which consists of the Chinese and the English translation of news stories and editorials from major news agencies and newspapers in mainland China. The parallel text that was used for training contains 12.0M Chinese characters and 9.7M English words in total. A small portion of parallel text(132K Chinese characters, 106K English words) from the same corpus was manually annotated to be used as the test data, hereafter referred to as PTtest_CN and PTtest_EN.

The comparable text were extracted from the 1995 – 2001 Xinhua sections of Chinese Gigaword Third Edition[3] and English Gigaword Third Edition[4], using the lexicon based content matching algorithm. A total of 15,133 document pairs, or 5.8M Chinese characters, 2.9M English words were extracted by this method.

4.2. The Experiments

The co-training algorithm was run on the parallel text for six iterations. The Chinese tagger and the English tagger at the end of each iteration were then tested on the ACE test data and the PTtest data.

The same experiments were also run on the comparable text, and the Chinese tagger and the English tagger at the end of each iteration were tested on the ACE test data.

5. Results

Figure 2 to 7 illustrate the precision, recall and f-measure of the co-trained taggers on different test sets, where *PT* stands for parallel text, and *CT* for comparable text. Note all f-measures improve significantly before deteriorating after the third iteration. The deterioration is caused by propagating noises coming from labeling and projection errors during the co-training. The F-measures with comparable text degraded much faster than with

[2] FBIS data were made available to TIDES and GALE researchers by LDC, but not to general public.

[3] LDC catalogue number LDC2007T38

[4] LDC catalogue number LDC2007T07

parallel text, because name projection with comparable text is more difficult and noises are easier to find their way into the training data.

Table 1 shows co-trained Chinese/English named entity taggers' performance on the ACE test data and PTtest data. Because co-training using parallel text was run six iterations, there are six co-trained Chinese taggers and six English taggers. Due to the space limitation, the table only shows the best f-measure (column *BestF*) achieved by the six taggers.

The table clearly shows that co-trained taggers have great improvement over the baseline taggers. In addition, in this experiment, using comparable text achieved about the same result as using parallel text. Note that using comparable text showed significantly better result on ACE Chinese test data than using parallel text.

The test results on ACE test data show that co-training with parallel and comparable test can effectively enhance a name tagger's performance in the domain the tagger was originally trained on.

The test results on PTtest demonstrate that co-training with bilingual text can be used to adapt existing taggers to new domains.

Test Set	Parallel Text		Comparable Text	
	Baseline	BestF	Baseline	BestF
ACE Chinese	80.45%	83.43%	80.45%	84.12%
ACE English	86.79%	88.55%	86.79%	88.29%
PTtest Chinese	84.89%	88.65%	NA	NA
PTtest English	81.55%	85.43%	NA	NA

Table 1 F-measures of co-trained taggers on test sets; BestF indicates the best F-measure co-trained taggers achieved

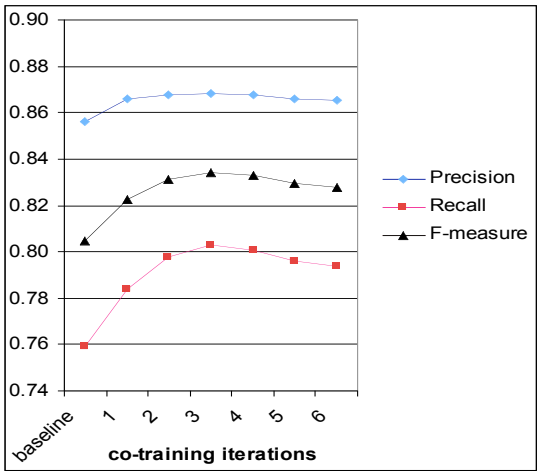

Figure 2 PT models on ACE Chinese test data

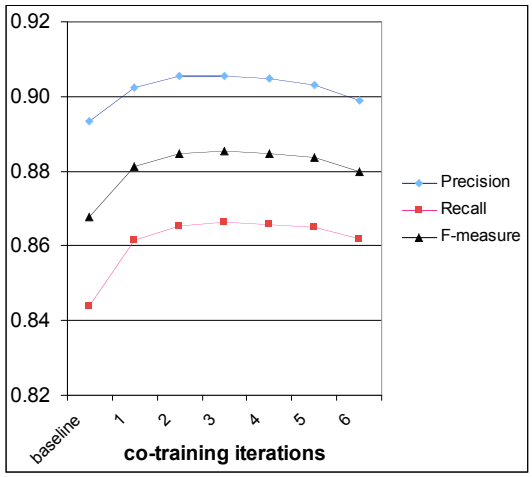

Figure 3 PT models on ACE English test data

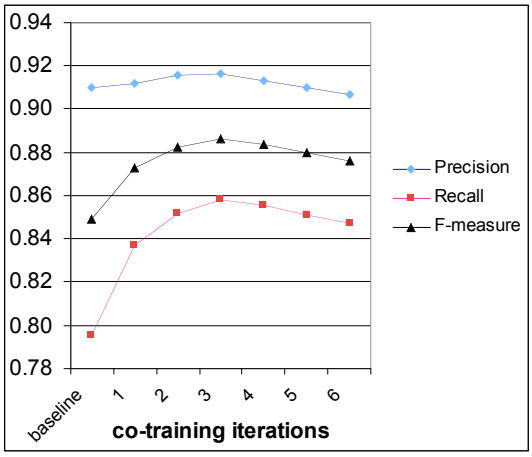

Figure 4 PT models on Chinese PTtest data

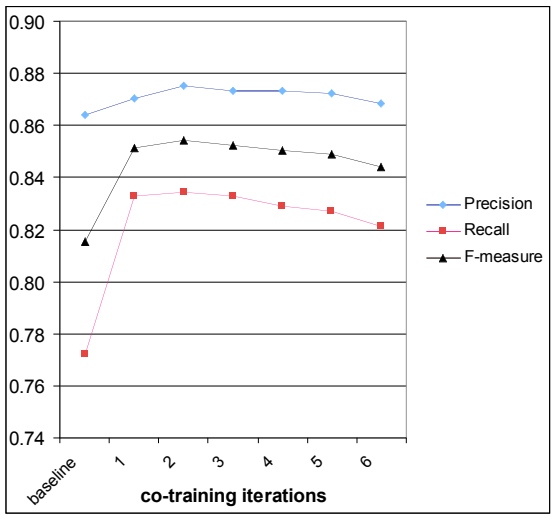

Figure 5 PT models on English PTtest data

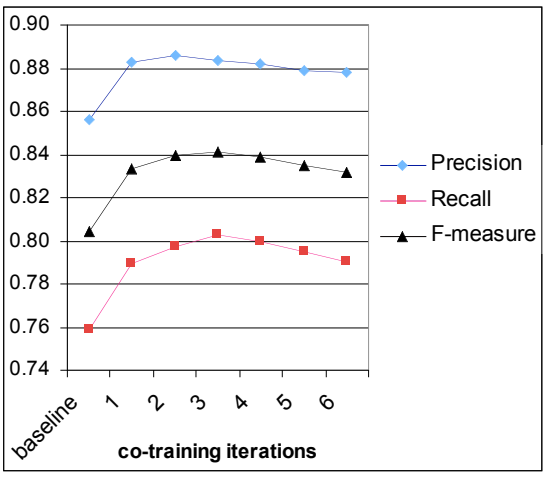

Figure 6 CT models on ACE Chinese test data

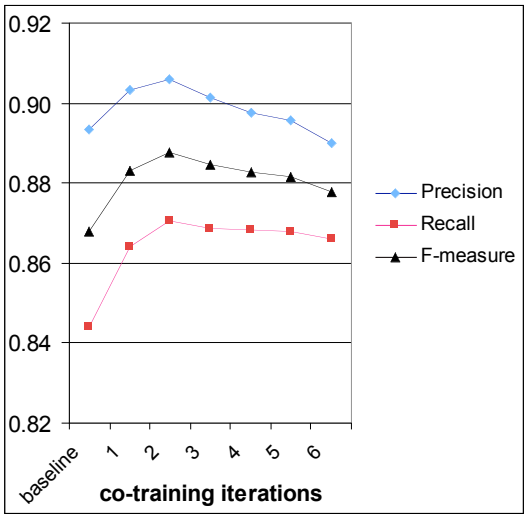

Figure 7 CT models on ACE English test data

6. Conclusion

We have demonstrated that applying co-training on unlabeled bilingual data can improve current state-of-the-art NE taggers, and adapt existing taggers to new domains. Together with entity alignment, we can extend our method from parallel text to comparable text, which has a much greater availability in many domains.

The co-training and entity alignment algorithm we presented have several advantages over previous approaches – the same algorithm can be applied on comparable text; the amount of data required to make the algorithm work is less than word alignment-based approaches; the algorithm can improve NE taggers of both sides of the bilingual text.

The improved tagger performance lays a solid foundation for future works on named entity alignment.

7. References

AL-ONAIZAN, Y., CURIN, J., JAHR, M., KNIGHT, K., LAFFERTY, J., MELAMED, D., OCH, F., PURDY, D., SMITH, N. & YAROWSKY, D. (1999) Statistical Machine Translation. Final Report, JHU Summer Workshop.

BIKEL, D. M., SCHWARTZ, R. L. & WEISCHEDEL, R. M. (1999) An Algorithm that Learns What's in a Name. *Machine Learning,* vol. 34**,** issue 1-3, pp. 211-231.

BLUM, A. & MITCHELL, T. (1998) Combining Labeled and Unlabeled Data with Co-training. *In Proceedings of the Workshop on Computational Learning Theory,* pp. 92-100. Morgan Kaufmann Publishers.

BORTHWICK, A. (1999) A Maximum Entropy Approach to Named Entity Recognition. Ph.D dissertation, New York University.

CUCERZAN, S. & YAROWSKY, D. (1999) Language independent named entity recognition combining morphological and contextual evidence. *In Proceedings of 1999 Joint SIGDAT Conference on EMNLP and VLC,* pp. 90-99.

DODDINGTON, G., MITCHELL, A., PRZYBOCKI, M., RAMSHAW, L., STRASSEL, S. & WEISCHEDEL, R. (2004) Automatic Content Extraction (ACE) program - task definitions and performance measures. *In Proceedings of LREC 2004: Fourth International Conference on Language Resources and Evaluation,* pp. 837-840.

GRISHMAN, R. (1995) The NYU System for MUC-6 or Where's the Syntax? *In Proceedings of the MUC-6 workshop,* pp. 167-175. Washington.

HWA, R., RESNIK, P., WEINBERG, A., CABEZAS, C. & KOLAK, O. (2005) Bootstrapping Parsers via Syntactic Projection across Parallel Texts. *Journal of Natural Language Engineering, special issue on parallel text,* 11:3**,** pp. 311-325.

LI, W. & MCCALLUM, A. (2003) Rapid Development of Hindi Named Entity Recognition using Conditional Random Fields and Feature Induction. *In Proceedings of ACM Transactions on Asian Language Information Processing, 2003,* pp. 290-294.

KOEHN, P., HOANG, H., BIRCH, A., CALLISON-BURCH, C., FEDERICO, M., BERTOLDI, N., COWAN, B., SHEN, W., MORAN, C., ZENS, R., DYER, C., BOJAR, O., CONSTANTIN, A. & HERBST, E. (2007) Moses: Open Source Toolkit for Statistical Machine Translation. *In Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session,* pp. 177-180. Prague, Czech Republic.

YAROWSKY, D., NGAI, G. & WICENTOWSKI, R. (2001) Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora. *In Proceedings of HLT 2001, First International Conference on Human Language Technology Research,* pp. 161-168.

Using Product Review Sites for Automatic Generation of Domain Resources for Sentiment Analysis: Case Studies

Eugenie Giesbrecht

FZI Forschungszentrum Informatik at the University of Karlsruhe,
Haid-und-Neu-Str. 10-14, 76131 Karlsruhe, Germany
giesbrecht@fzi.de

Abstract

In this work, we explore the usage of *Amazon Web Services* to automatically acquire domain sentiment resources for consumer products in English and German. We empirically evaluate the automatically gained corpora for the tasks of sentiment classification and domain-oriented sentiment lexicon extraction. The preliminary results are encouraging.

1. Motivation

There is an increasing availability of opinion and sentiment rich resources on the World Wide Web. As a consequence, even more decisions are being taken based on the information available online. Whether we speak of enterprise decision makers or casual users, 81% of internet users utilized internet at least once to find information about some products. In 73-87% cases, online available reviews influenced people's purchase intentions, so that they were ready to pay twice as much for the product that was rated higher (Pang and Lee, 2008). Thus, monitoring sentiments online has become an indispensable exercise of both private internet users and business analysts.

In NLP research, in return, sentiment analysis has been extensively studied in the recent years. Among the pioneering papers in this area are those of (Pang et al., 2002; Turney, 2002) on sentiment classification. *Sentiment classification* is a special kind of text categorization where the task is to classify text snippets according to their sentiment towards a given subject.

Due to the intrinsic complexity of the task, it is in practice usually reduced to identifying the authors positive of negative attitude towards the topic, i.e. the polarity of a text segment.

It has been shown in the recent literature that sentiment classification is a very domain specific task (Aue and Gamon, 2005; Li and Zong, 2008), since sentiments in different domains can be expressed in rather different ways. Natural language resources for automatic analysis all tend to be confronted sooner or later with the problem of domain dependency, i.e. the applications built with statistical models are just as good as the corpora on which the models were trained, and usually they perform well in the domains on which they were trained. Applications building upon the available thesauri or gazetteers, in the similar way, can be just as good as the underlying thesauri.

This is a well-known problem in computational linguistics since the spread of the World Wide Web. This problem is usually addressed in two ways, either by training existing classifiers on new domain data, or by adapting existing classifiers for new domains. While the former is generally considered unrealistic as manually annotated resources for new domains are unavailable; the latter received a growing attention in recent NLP research. In spite of quite a number of suggested domain adaptation algorithms (Daum and Marcu, 2006; Blitzer et al., 2007a; Dredze and Crammer, 2008), any of those would only benefit if more annotated domain resources were available.

Acquisition of domain as well as language specific resources is generally considered a time-consuming and expensive exercise. While this is true for many NLP tasks, there are lots of user-generated contents that can be employed for acquisition of domain specific sentiment resources as well as product or topic centered resources where manually created content is available.

We are not the first to use product reviews for sentiment analysis research. In 2004, Hu and Liu (2004a) annotated customer reviews of 5 products taken from *Amazon.com* for the experiments on sentiment summarization as well as feature extraction (Hu and Liu, 2004b). In 2007, Blitzer et al. (2007b) collected a multi-domain sentiment dataset[1] from *Amazon.com* for 4 product types: *kitchen*, *books*, *DVDs*, and *electronics*, in order to evaluate their domain adaptation method[2]. This sentiment dataset has been used in a number of further works on domain adaptation in sentiment classification (Mansour et al., 2008; Li and Zong, 2008).

Product review sites also gain popularity in traditional *text classification* research. For example, Ifrim et al. (2008) use movie reviews from *IMDB*[3] to learn the movie genre from the short plot description associated with each movie and a dataset of editorial reviews of books from *Amazon* for the task of book reviews classification by genre.

Using product reviews is in the meantime the standard in the task of *aspect-based sentiment summarization* where the metadata of the reviews are used for the extraction of product specific features. Branavan et al. (2008) crawl reviews for restaurants and cell phones with attached lists of keyphrases from *Epinions.com* website to show that a joint model of text and user annotations can benefit extractive

This work is supported by German "Federal Ministry of Economics" (BMWi) under the project Theseus (number 01MQ07019).

[1] http://www.cs.jhu.edu/ mdredze/datasets/sentiment/index2.html

[2] Later, the original dataset was expanded to more domains: http://www.cs.jhu.edu/ mdredze/datasets/sentiment/

[3] http://www.imdb.com/

summarization. Titov and McDonald (2008) refer to hotel reviews from *TripAdvisor.com* to leverage existing aspect ratings to learn mappings from text to aspects and extract fragments of text discussing these aspects without the need of annotated data. Meng and Wang (2009) mine Chinese product specifications from ZOL[4] to automatically extract product features from user reviews and to generate a review summary. Du et al. (2010) use three domain-specific datasets - hotel (from www.ctrip.com), electronics (from detail.zol.com.cn) and stock (from blog.sohu.com/stock) - to construct domain sentiment lexicons for Chinese.

To the best of our knowledge, we are not aware of similar efforts for languages other than English and Chinese.

Movie, products, hotel reviews, topic and mood labels in blogs and many other Web 2.0 resources offer a unique source of semi-structured information that can be employed for numerous NLP tasks. Two of those, namely domain corpora construction and domain lexicon extraction, are discussed in this paper. The main contribution of this work is the study of suitability of online product review sites, like *Amazon.com* for generation of multilingual domain resources.

The rest of the paper is organized as follows. We start by providing an overview of the automatically acquired domain corpora in Section 2. Section 3. describes two task-based evaluations of those corpora - for sentiment classification and domain-oriented sentiment lexicon extraction. In the end, we conclude and point out avenues for further research in Section 4.

2. Domain Resources Construction

Our goal is to automatically build domain specific corpora for consumer products and specifically for sentiment analysis. For this, we deploy *Amazon Web Services*[5] to get a direct access to reviews for specific product groups. Each *Amazon* review consists of a rating (0-5 stars) and a number of further metadata (e.g. a reviewer's name, a review date, etc). The latter are irrelevant for our purposes, so we ignore them for the moment of being and collect only the review content for 3 selected subcategories ordered by stars. In contrast to the available sentiment collections that have been collected for rather coarse defined domains, we crawl data for more fine-grained categories, i.e. instead of building a general *electronics* corpus as in Blitzer et al. (2007a), we gather reviews for *cell phones*, *digital cameras* and *mp3 players* which are all subcategories of *electronics*.

We collect 5-star and 1- & 2-star reviews of the *top sellers* within given categories as representative for positive and negative sentiments. We discard the rest as their polarity can be mixed. Table 1 shows the number of the overall collected reviews[6] for the purposes of this study.

In order to evaluate the quality of the resulting domain corpora, we employ the latter in two typical use cases: for sentiment classification and for gaining domain sentiment lexicons.

[4] www.zol.com.cn

[5] http://aws.amazon.com/

[6] The size of the dataset was primarily limited by the number of reviews we were able to crawl within available time.

Domain	Language	Polarity	N. Reviews
digital cameras	English	neg	257
		pos	810
	German	neg	67
		pos	637
cell phones	English	neg	141
		pos	429
	German	neg	337
		pos	593
mp3 player	English	neg	438
		pos	984
	German	neg	165
		pos	781

Table 1: Number of the crawled reviews pro domain used for the case studies

3. Case Studies

3.1. Sentiment Classification

We evaluate the above described automatically constructed corpora on a sentiment classification task. Sentiment classification is defined here as the assignment of texts into either positive or negative groups. In the first step, we test a straightforward approach, motivated by the assumption that the content of reviews is per se subjective and contains sentiments.

For classification, we make use of a reimplementation of the basic sentiment classifiers described in Pang and Lee (2004) using Dynamic Language Modeling Classifier available from LingPipe[7] with default settings as described in the Lingpipe's sentiment analysis tutorial[8][9]. The latter is a language modeling classifier with training based on a multivariate estimator for the category distribution and dynamic language models for the per-category character sequence estimators.

We test the following six settings for two languages - English and German:

1. train and test on *cell phones*
2. train and test on *digital cameras*
3. train and test on *mp3 player*
4. train on *cell phones* and test on *digital cameras*
5. train on *cell phones* and evaluate on *mp3 players*
6. train and test on a mix of all three domains

Table 2 shows accuracies[10] for German for two *training / test set* cutoffs. Interestingly, using a bigger training set does not automatically bring about better sentiment classification accuracy in German. This is presumably due to

[7] http://alias-i.com/lingpipe/

[8] http://alias-i.com/lingpipe/demos/tutorial/sentiment/readme.html

[9] As our objective here is not the improvement of the algorithm but assessment of the corpus quality, we make use of the available classifiers for our experiments

[10] Accuracy is defined as the number of correctly classified test texts

the unbalanced corpora we used for the experiments. Further evaluation has to be done here. However, even with small and unbalanced corpora, the achieved accuracies are all at or above reported state-of-the-art levels in sentiment classification for English (cf. Aue and Gamon (2005), Li and Zong (2008))[11]. Furthermore, we are not aware of any reported results for this task for German.

In the second step, we evaluate the extracted domain corpora for English with two classification paradigms. Additionally to the above-described setup, we test the hierarchical classification technique described in Pang and Lee (2004). With hierarchical classification, *subjective sentences* are identified first. Those are then forwarded to the polarity classifier.

Table 3 shows the resulting accuracies for English. Here, using more training examples is consistently and significantly better than using less training. However, preliminary experimentation with automatic filtering of subjective sentences before classifying polarity of the text did not cause any significant accuracy improvements, in some cases even accuracy decrease.

3.2. Domain (Sentiment Polarity) Lexicon Extraction

Automatic construction of domain-oriented sentiment lexicons is a further important task in sentiment analysis (Du et al., 2010).

In order to evaluate the quality of the automatically extracted domain sentiment corpora (see Section 2.) for this task, we make use of the Java Automatic Term Recognition Toolkit (JATR)[12] - a toolkit for developing and experimenting with automatic term recognition algorithms (Zhang et al., 2008).

The JATR toolkit includes 5 state-of-the-art term extraction algorithms: TF-IDF, C-Value (Frantzi and Ananiadou, 1999), Weirdness (Ahmad et al., 1999), Glossex (Kozakov et al., 2004), and Termex (Sclano and Velardi, 2007). While the first two algorithms employ only the statistics of the terms within the given text collection, the rest three methods exploit term statistics in the given corpus in respect to the reference corpus, i.e. they aim at taking into consideration the uniqueness of specific domain corpus terminology.

We arbitrarily choose the *cell phones* corpus and start our experiments with *Weirdness* algorithm. It compares term frequencies in the target and the reference corpus. The British National Corpus *BNC*[13] corpus statistics is provided within JATR package so that one can easily use it as a reference corpus of general English language. Table 4 show top terms extracted from negative and positive reviews respectively in contrast to *BNC*.

Manual investigation of the *top 30 terms* extracted by *Weirdness* algorithm using the *BNC* as a reference corpus reveals a lot of overlap between positive and negative reviews. Consequently, *top noun terms* have turned out to be not very distinctive for polarity, but are still good candidates for the domain *cell phone* lexicon. If we further differentiate between the extracted nouns[14] vs adjectives, we can still observe a certain *polarity tendency* (see Table 4).

Thus, using the *BNC* as a reference corpus seems to work well for gaining domain lexicons (cf. Table 5). However, it appears to be not quite enough for domain sentiment polarity lexicons. In the next step, we extract terminology from the corpus of positive reviews using the corpus of negative reviews as a reference corpus and vice versa. Tables 6 and 7 demonstrate top 10 terms and top 10 phrases respectively, for both English and German. Here, the resulting top terminology is very sentiment biased, especially phrases, i.e. multiword units, seem to be more adequate for domain-oriented sentiment lexicon.

Thus, term extraction algorithms per se are not sufficient for compiling domain sentiment terminology using general purpose corpora as a reference. For this, we suggest a simple and intuitive alternative, namely to use the automatically crawled corpora from the same domain (*here: cell phones*) but of the opposite polarity. E.g. using *Weirdness* algorithm, the lexicon entry for *great camera* appears at position 21 (using the corpus of opposite polarity) vs at position 867 (using the BNC as a reference corpus).

4. Conclusions and Outlook

In this paper, we exploit the applicability of user-generated reviews as a source of automatic construction of domain sentiment corpora and domain (sentiment) lexicons.

The preliminary results of the task-based evaluation of the domain corpora automatically extracted with the help of *Amazon Web Services* are very encouraging. There is still a lot to explore. One of the big questions is further assessment of the quality of automatically gained corpora, as there is limited control over the contents of the resulting corpus. Consequently, further post evaluation is presumably needed to estimate the actual corpus content (Ferraresi et al., 2008). There are plenty of opportunities to employ such automatically generated domain corpora in sentiment analysis applications.

Though we concentrate here on domain corpora for sentiment analysis for consumer products, one can surely benefit from available user-generated reviews in any other domains and languages where those are available.

Acknowledgements. Special thanks to the anonymous reviewers for their constructive critique.

5. References

Khurshid Ahmad, Lee Gillam, and Lena Tostevin. 1999. University of surrey participation in TREC8: Weirdness indexing for logical document extrapolation and retrieval (wilder). In *TREC*.

Anthony Aue and Michael Gamon. 2005. Customizing sentiment classifiers to new domains: A case study. In *RANLP-05, the International Conference on Recent Advances in Natural Language Processing*.

[11] However, to the best of our knowledge, no comparable evaluation has been made for the currently used fine-grained level of domain categories.

[12] http://www.dcs.shef.ac.uk/ ziqizhang/project

[13] http://www.natcorp.ox.ac.uk/

[14] Names of companies - such as Nokia, Sony, etc - were not considered.

	small training set (number of training/test texts)	bigger training set
cell phones	(205/725) 86 %	(725/205) 88%
mp3 player	(202/744) 84%	(744/202) 82%
digital cameras	(212/734) 85%	(734/212) 79%
cell phones $->$ mp3 players		(930/946) 88,5%
mp3 players $->$ cell phones		(946/930) 67%
cell phones $->$ digital camera		(930/704) 87%
digital camera $->$ cell phones		(704/930) 64,2%
cross-domain	(546/2034) 86,2%	(2034/546) 84,6%

Table 2: Classifier Accuracies for German

	small training	bigger training	using subj filter (small)	using subj filter (big)
cell phones	(113/457) 76,1 %	(457/113) 81,4%	(113/457) 75,9%	(457/113) 81,4%
mp3 player	(286/1136) 75%	(1136/286) 81%	(286/1136) 75,3%	(1136/286) 82,2%
digital cameras	(211/856) 78,3%	(856/211) 86,7%	(211/856) 78,4%	(856/211) 88,2%
cell phones $->$ mp3 players		(570/1422) 75,4%		(570/1422) 75%
mp3 players $->$ cell phones		(1422/570) 80%		(1422/570) 77,5%
cell phones $->$ cameras		(570/1067) 74,5%		(570/1067) 78,4%
cameras $->$ cell phones		(1067/570) 79%		(1067/570) 78,6%
cross-domain	(610/ 2449) 77,5%	(2449/610) 84%		

Table 3: Classifier Accuracies for English

John Blitzer, Mark Dredze, and Fernando Pereira. 2007a. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic, June. Association for Computational Linguistics.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007b. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the Association for Computational Linguistics (ACL)*.

S.R.K. Branavan, Harr Chen, Jacob Eisenstein, and Regina Barzilay. 2008. Learning document-level semantic properties from free-text annotations. In *Proceedings of ACL-08: HLT*, pages 263–271, Columbus, Ohio, June. Association for Computational Linguistics.

Hal Daum and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. *J. Artif. Intell. Res. (JAIR)*, 26:101–126.

Mark Dredze and Koby Crammer. 2008. Online methods for multi-domain learning and adaptation. In *EMNLP*, pages 689–697. ACL.

Weifu Du, Songbo Tan, Xueqi Cheng, and Xiaochun Yun. 2010. Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon. In Brian D. Davison, Torsten Suel, Nick Craswell, and Bing Liu, editors, *WSDM*, pages 111–120. ACM.

A. Ferraresi, E. Zanchetta, M. Baroni, and S. Bernardini. 2008. Introducing and evaluating ukWaC, a very large Web-derived corpus of English. In *Proceedings of the WAC4 Workshop at LREC 2008*.

Katerina T. Frantzi and Sophia Ananiadou. 1999. The C/NC value domain independent method for multi-word term extraction. *Journal of Natural Language Processing*, 6(3):145–180.

Minqing Hu and Bing Liu. 2004a. Mining and summarizing customer reviews. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, New York, NY, USA. ACM.

Minqing Hu and Bing Liu. 2004b. Mining opinion features in customer reviews. In *AAAI'04: Proceedings of the 19th national conference on Artifical intelligence*, pages 755–760. AAAI Press / The MIT Press.

Georgiana Ifrim, Gökhan Bakir, and Gerhard Weikum. 2008. Fast logistic regression for text categorization with variable-length n-grams. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 354–362, New York, NY, USA. ACM.

L. Kozakov, Y. Park, T.-H Fin, Y. Drissi, Y. N. Doganata, and T. Confino. 2004. Glossary extraction and utilization in the information search and delivery system for IBM Technical Support. *IBM Systems Journal*, 43(3).

Shoushan Li and Chengqing Zong. 2008. Multi-domain sentiment classification. In *HLT '08: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*, pages 257–260, Morristown, NJ, USA. Association for Computational Linguistics.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. 2008. Domain adaptation with multiple sources. In *NIPS*, pages 1041–1048.

Xinfan Meng and Houfeng Wang. 2009. Mining user reviews: from specification to summarization. In *ACL-IJCNLP '09: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 177–180, Morristown, NJ, USA. Association for Computational Linguistics.

1*/2*: nouns	5* reviews: nouns	1*/2*: adjectives	5*: adjectives
browser	touchscreen	chunkiest	resistive
bluetooth	bluetooth	mistyping	capacitive
firmware	firmware	swiping	unlocked
wtf	usb	resisitive	responsive
traveler	headset	bricking	resistive
messaging	email	incompetent	favorite
helpdesk	wrt	unlocked	capacitive
browser	qwerty	capacitive	customizable

Table 4: Top Nouns vs Top Adjectives with Weirdness Algorithm using BNC as reference corpus

cell phones (Weirdness)	cell phones (TermEx)	cell phones (GlossEx)	cameras (Weirdness)	mp3-player (Weirdness)
touchscreen	game	touchscreen	slr	usps
bluetooth	pic	bluetooth	mov	nano
firmware	webpage	usb	zoom	flac
usb	contact	customizable	camera	belkin
customizable	minute	browser	usb	usb
bluetooth headset	feature	wtf	graininess	fuse
email	instruction	dialer	blurry	sync
wrt	app	qwerty	nimh	jbl
resistive touchscreen	a-gps	sync	hd camera	favorite
messaging	ton	traveler	reba	nano-battery
headset	expectation	inbox	energizer	pedometer
qwerty keypad	document	download	lens	pda
browser	aplicacione	firmware	lcd	usb charger

Table 5: Top 10 terms with 3 different JATR algorithms for *cell phones* and with *Weirdness* for *cameras* and *mp3 players* using the *BNC* as a reference corpus

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*, pages 271–278.

Bo Pang and Lillian Lee. 2008. *Opinion Mining and Sentiment Analysis*. Now Publishers Inc.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of EMNLP-02, the Conference on Empirical Methods in Natural Language Processing*, pages 79–86, Philadelphia, US. Association for Computational Linguistics.

Francesco Sclano and Paola Velardi. 2007. TermExtractor: a web application to learn the shared terminology of emergent web communities. In *Proc. of the 3rd International Conference on Interoperability for Enterprise Software and Applications (I-ESA 2007)*.

Ivan Titov and Ryan McDonald. 2008. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of ACL-08: HLT*, pages 308–316, Columbus, Ohio, June. Association for Computational Linguistics.

P Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings 40th Annual Meeting of the Association for Computational Linguistics*, ACL, pages 417–424.

Ziqi Zhang, José Iria, Christopher Brewster, and Fabio Ciravegna. 2008. A comparative evaluation of term recognition algorithms. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC08)*, Marrakech, Morocco.

terms (EN)	phrases (EN)	terms (DE)	phrases (DE)
t plan	great buy	zufrieden	sehr zufrieden
t network	flash player	htc	super schnell
easy	great price	top	voll zufrieden
buy	great camera	sprachqualität	total zufrieden
life	great keyboard	schnell	wirklich fr meine
plan	good buy	telefonieren	top zufrieden
wait	camera flash	super	super gert
t store	great smartphone	gert	sprachqualitt sehr gut
love	network set	bedienung	allem sehr zufrieden mit
great	great deal	begeistert	wikrlich schn

Table 6: Top 10 terms and phrases for 5* reviews for cell phones, using 1* and 2* reviews as a reference corpus

terms (EN)	phrases (EN)	terms (DE)	phrases (DE)
click	no warranty	keine	gibt keine
usage	hard use	schreiben	immer wieder
fact	no click	wieder	keine bedienung
no	poor replacement	tastensperre	es gibt keinen
process	hard plastic	sterne	mehr geld
today	working fine	sieht	absolut schlecht
wait	cheap plastic	geld	schon wieder
use	bad buy	fazit	und wieder

Table 7: Top terms and phrases for 1*/2* reviews for cell phones, using 5* reviews as a reference corpus

Using machine learning to perform automatic term recognition

Jody Foo, Magnus Merkel

Department of Computer and Information Science,
Linköping university
jodfo@ida.liu.se, magme@ida.liu.se

Abstract

In this paper a machine learning approach is applied to Automatic Term Recognition (ATR). Similar approaches have been successfully used in Automatic Keyword Extraction (AKE). Using a dataset consisting of Swedish patent texts and validated terms belonging to these texts, unigrams and bigrams are extracted and annotated with linguistic and statistical feature values. Experiments using a varying ratio between positive and negative examples in the training data are conducted using the annotated n-grams. The results indicate that a machine learning approach is viable for ATR. Furthermore, a machine learning approach for bilingual ATR is discussed. Preliminary analysis however indicate that some modifications have to be made to apply the monolingual machine learning approach to a bilingual context.

1. Background

Term extraction, or more specifically, Automatic Term Recognition (ATR) is a field in language technology that involves "extraction of technical terms from domain-specific language corpora." (Zhang et al., 2008). One area of application of ATR systems is the field of terminology where one task is to define concepts and decide which terms should be attached to them. Another area of application is creating domain specific dictionaries for use in for example machine translation (Merkel et al., 2009).

Closely related fields to ATR are Automatic Keyword/Keyphrase extraction, (AKE) (Turney, 2000; Hulth, 2004), and Automatic Index Generation (AIG) and Information retrieval (IR) in general. Common to these fields are the following components.

- text analysis
- selecting/filtering term candidates
- ranking term candidates

The first component involves the process of analyzing the source text(s) and attaching feature value pairs to each token in the text. The second component involves the process of selecting single or multi-word-units from a text. The selection process may be iterative, e.g. it may consist of selection and filtering in several iterations. In most cases, filtering relies on the feature-value pairs previously attached. The third component involves using one or several metrics to rank term candidates.

2. Current research

In order to decide what should be considered a term, it is necessary to analyze and attach some kind of information to the objects of consideration. In general, the kind of information produced by a method of analysis falls into one of two categories. Either it is 1) *statistical/distributional* or 2) *linguistic*. Statistical information is based on statistical analysis, e.g. word counts, probabilities, mutual information etc. Distributional properties can also be grouped into this category, e.g. corpora frequency comparisons. Linguistic information is based on linguistic analysis e.g. part of speech, semantics etc.

Although research in the fields of ATR, AKE and IR have common components, it is interesting to note that there seems to be little interaction between researchers within these fields, especially between ATR and AKE. It is also interesting to note that early ATR research (Bourigault, 1992; Ananiadou, 1994) relied on linguistic analysis, where as early work in Information retrieval relied on statistical measures (Salton and McGill, 1986). Current ATR research has however moved towards using statistical measures (Zhang et al., 2008) and current AKE research has moved towards using more linguistic analysis (Hulth, 2004). These movements, though moving towards the same center of conversion, seem to be independent of each other.

2.1. Statistical analysis

The most basic statistical measure is the frequency count, which has can be refined by normalizing, and more commonly, by adding a distributional component as with the *tf-idf* measure (Salton and McGill, 1986). In ATR a distinction between "unithood" and "termness". Unithood measures are used to determine collocation strength between units when dealing with terms that consist of more than one word. Termness measures indicate how associated to the domain a term is. Typical unithood measures are mutual information (Daille, 1994) and log-likelihood (Cohen, 1995a). Statistical measures such as the C-value/NC-value (Frantzi et al., 1998) integrate termness unithood. The C-value part of this metric is tailored to recognize termness. The NC-value takes context into account and thereby improves multiword unit recognition.

2.2. Linguistic analysis

Early term recognition systems such as (Bourigault, 1992; Ananiadou, 1994) used part-of-speech and chunking to facilitate the use part-of-speech patterns to recognize term candidates. Frameworks relying solely on linguistic analysis use hand-crafted rules such as {`DET A N`} to recognize term candidates.

2.3. Term candidate selection

The term candidate selection process is the process of selecting which of the extracted terms should be passed on to e.g. a domain expert for validation. To do this, the initial set of possible term candidates is truncated into a smaller set using a metric that measures termness. The better the termness value, the more likely it is that the extracted term candidate will be accepted by a domain-expert.

2.4. Machine learning

To the authors' knowledge, machine learning has not been applied to ATR. However, it is a common approach within AKE (Turney, 2000; Hulth, 2003; Hulth, 2004). (Turney, 2000) performed experiments using C4.5 (Quinlan, 1993) among others. (Hulth, 2004) used a rule induction system called Compumine[1]. Turney (2000) used 10 features in his C4.5 experiments, most of them linguistically uninformed, such as number of words in the phrase, frequency of the phrase and relative length of phrase. Three of the features used heuristics to determine whether the phrase was a proper noun, ended by an adjective or contained a common verb. The experiments carried out by (Hulth, 2003) used the same features as (Frank et al., 1999) but also add a string containing part-of-speech tags.

3. Problem specification

As machine learning has been successfully applied to AKE, it is relevant to examine its efficiency when applied to ATR. Though the extraction processes of ATR and AKE are similar, the two tasks are different when looking at the expected output. The task of AKE is to output a relatively short list of keywords/keyphrases that describe a document. This size of this list is between 5 and 15 keywords/keyphrases long. In ATR there is no limit on how many terms are extracted. One problem when applying machine learning techniques is that the composition of the training data can have a drastic effect on precision and recall. Given that one property of terms in documents is that they make up only a small percentage of the total amount of tokens, it is important to examine how this property effects training results. Performing experiments with different positive/negative example ratios will add two important pieces of knowledge for future research.

1. Is a machine learning approach feasible for ATR?
2. What should the ratio be between positive and negative examples in the training data for future experiments?

In the remaining part of this paper we will describe such experiments, their results and possible directions for future research.

4. Method

The method examined in this paper uses Ripper (Cohen, 1995b) a rule induction learning system that produces human readable rules. Using a rule producing machine learning algorithm has the advantage of making it possible for humans to read the rules and try to understand what a mechanical algorithm deems as important features of terms. Produced rules can also be documented and used in other systems.

4.1. Evaluation

Common measures used to evaluate term extraction results are precision, recall and f-score. However, even though the metrics may be the same, applications of the metrics differ between ATR and AKE. As mentioned, in the task of keyword extraction the output is a small list of 5-15 keywords and the performance is measured over this list. Turney (2000) e.g. measures the precision of the first 5, 7, 9, 11, 13, and 15 phrases. In contrast, the length of the output from a ATR system does not have a formal limit, which means that precision and recall numbers cannot be compared between the two tasks.

Zhang et al. (2008) discusses several possible evaluation metrics, noting that many the evaluation metrics used "only measure precision but not recall" and that "they evaluate only a subset of the output". With the scenario of using ATR output to produce some kind of terminology resource in mind, e.g. as in Merkel et al. (2009), recall is of utmost importance, outranking precision. The reason is that when post-processing the list of extracted terms, it *is possible* to increase the precision of the final result by manually removing incorrect term candidates but, it *is impossible* to increase the recall above the recall of the original list.

Evaluating a subset of the output is not an option in our case as a the design of our experiment does not output a ranked list of term candidates. On the other hand, a subset evaluation method might not be as relevant in ATR as it is in AKE since this evaluation method is precision-oriented, rather than recall-oriented.

4.2. Dataset

The dataset used was provided by Fodina Language Technology and consists of Swedish patent texts grouped by IPC classes. A set of manually validated terms is also provided for each group of patent texts in the data set. The experiments in this paper were run on the smaller A42B subset (*hats; head coverings*) and the larger A61G subset (*transport, personal conveyances, or accommodation specially adapted for patients or disabled persons; operating tables or chairs; chairs for dentistry; funeral devices*). The composition of these data sets is described in tables 1 and 2. For this data set, the term segments refers to a line in the corpus text file which in most cases is a full sentence, but a segment can also be heading or a caption. As can be seen in table 2 there are very few two-word terms, and no three-word terms. This is due to the corpus being in Swedish language where compound nouns are frequently used. For example terms such as "file manager" or "file manager window" would both be a single word terms in Swedish: "filhanterare" and "filhanterarfönster").

As the data was made available as a large concatenated text file, document/document collection distributional measures such as *tf-idf* are not possible to calculate. The terms accompanying the patent documents had been previously extracted and were manually validated by domain experts (Merkel et al., 2009).

[1] http://www.compumine.com

Corpus statistics	A42B	A61G
Number of tokens	71761	302027
Number of segments	2929	12684
Number of terms (types)	579	1260

Table 1: Overview of the A42B and A61G document collections.

Term length	A42B	A61G
1 word terms	570	1240
2 word terms	9	20
3 word terms	0	0

Table 2: Composition of the validated term lists. Lack of 2 and 3 word terms is due to use of compounds in the Swedish language.

4.2.1. Feature selection

Based on previous research in ATR, we have chosen to use linguistic features as well as statistical features. Linguistic features were obtained using the commercial tagger Connexor Machinese Syntax[2]. A detailed explanation of these can also be found in (Ahrenberg, 2007). A normalized frequency count was also included. By using a normalized frequency count, the generated rules may also be generalized to other corpora. Furthermore based on previous studies conducted by Foo, a statistical language model of a tokenized general corpus was created and statistics derived from this language model were used as additional features. In the experiments conducted here, the PAROLE corpus[3] was used to build the language model. The motivation behind using a general language language model is to be able to capture how common a word or phrase is in non-domain specific text. All in all, 10 different features were used, as seen in table 3.

Feature	Description
POS	part-of-speech tag
msd	morpho-syntactic description
func	grammatical function
sem	semantic information
nfreq	normalized n-gram frequency in text
zeroprobs	number of tokens with zero probability in given the language model
logprob	the logistic probability value, ignoring unknown words and tokens
ppl1	the geometric average of 1/probability of each token, i.e. perplexity
ppl2	the average perplexity per word

Table 3: List of features used to annotate the examples used in training and test data.

4.3. Preprocessing

The patent texts were provided as one document per subclass, i.e. one document for subclass A42B and one document for subclass A61B. These documents were tagged using Connexor Machinese Syntax and then n-gram extraction was performed which created separate files for each n-gram length while creating the n-gram files. Frequency counts were also done. The files were then annotated with statistics using the language model and finally each n-gram was annotated as a term or a non-term based on the validated term lists.

The result after all preprocessing had been completed was one feature-annotated file for each n-gram length. This file only contains unique examples, with respect to words and linguistic information, i.e. the word "speed" may exist in two unigram rows, but in one row it is tagged as a noun and in the second row it is tagged as a verb.

4.4. Experiments

The key variable in the experiments described in this paper is the positive/negative example ratio used in the training data. We chose five different ratios: 10/90, 30/70, 50/50, 80/20, and 90/10 where the first number is the number of positive examples.

We chose to create separate systems for each n-gram length in our experiments, i.e. one system would create rules for unigrams and another for bigrams. Training and test data used a feature representation which treats tokens in e.g. a bigram as single entities. An alternative is to group tokens into one value. That is, we use a non-aggregated form, e.g. `BIGRAM="SMALL CAR" POS1=A POS2=N` rather than `TEXT="SMALL CAR" POS=A_N`.

In total, 10 experiments per corpus were run, totaling 20 experiments for both corpora. The original plan was to include trigrams, but since no three word terms were included in the term lists, only unigram and bigram experiments were conducted.

After annotating the extracted n-grams according to the process described in section 4.3., 10% randomly chosen example rows from each n-gram set was held back to be used as test data. As positive/negative example ratio is of interest, such properties of the unmodified training set and the test sets are presented in tables 4 and 5. Please note that the ratios reflect unique n-gram data and therefore does not represent an actual term/non-term token ratio of the documents.

The same test set was used for all experiments from within a n-gram group and corpus, e.g. the A42B unigram test set was used for all five unigram experiments conducted on the A42B corpus.

Ripper was run with the following settings for all experiments, `-a given -L 0.4`. The first option `-a given` means that Ripper is forced to use a specified class order. In practice this is a way to force Ripper to produce rules for the "positive" (term) class, even when there are more "negative" (non-term) examples in the training data. The second used option, `-L 0.4` sets the "loss" ratio to 0.4. The loss ratio is the ratio between the cost of a false negative compared to a false positive. A ratio of 0.4 as used in the experiments in this paper, provided a good recall progression in the experiments, i.e. it is possible to produce rules that achieve 100% recall without tipping the scale too much so that all rules sets produce 100% recall. All other settings were left to their default value. Though the set-

[2]http://www.connexor.eu/technology/machinese/machinesesyntax/
[3]http://spraakbanken.gu.se/parole/

tings might have an effect on precision and recall, the point of the experiments are not to find out how the best system performs.

5. Results

The results of the experiments are presented in tables 6 and 7. The general trend is the Ripper system produces rule sets that produce higher recall, the higher the positive/negative ratio is. However, as expected, the precision drops accordingly. The result tables list first unigram experiments then bigram experiments with varying positive/negative example ratios. For example experiment 1-10 refers to a unigram experiment with 10% positive examples in the training data. Experiment 2-30 would be a bigram experiment with 30% positive examples in the training data. The test data set to which the rules are applied are the held back test data described in tables 4 and 5.

As can be seen in the tables, the recall for many of the experiments is 100%. However, this does not mean that all existing terms in the corpus are among the extracted term candidates, it only says that 100% of the validated terms are present. Similarly, precision is relative to the validated term set. However, due to the nature of recall and precision, the real recall can only be equal or lower than the presented recall and the precision can only remain at its current level or improve given a more complete term list. Also, for our experiments, rows with 0 false negatives are the result of rulesets which simply classify all examples as terms. This makes sense from a precision point of view if training data shows that an overwhelming number of examples are terms.

5.1. Rules learned

As it would be unpractical to publish all learned rulesets in this paper, we will only include two sets, one for each n-gram length. These rules are presented in tables 8 and 9. The first rule in table 8 should be read as classify the example as a term (the 'yes' class) `IF` the example frequency (`freq`) is higher or equal to $2.7903e^{-5}$ and its part-of-speech tag is Noun (`n`). The rules are applied in order, i.e. for each example, try to apply a the first rule, if it is not applicable, use the next rule and so on. Interpreting the rules in 8 would produce the following:

- Nouns occurring more than X times are terms
- Nouns above a certain probability given the language model are terms
- All verbs are terms
- All singular nominals are terms
- All plural nominals are terms
- All adjectives are terms
- All singular genitive nouns are terms
- Perfect participles are terms

As can be seen, the rules learned by Ripper use both linguistic and statistical features.

6. Discussion

Comparing the results with e.g. (Zhang et al., 2008) is not possible as our experiment output is not ranked and can therefore not be pruned into a smaller set with higher precision. Including ranking metrics to the n-gram annotation is a good next step for future research. This would also mean that the machine learning algorithm can use this information as well in its learning process.

Besides using the learned rules to extract term candidates, it is also possible to use feature rich data, i.e. data with many levels of annotation, in combination with machine learning to discover new properties of terms. When it comes to finding rule patterns from huge amounts of data, a rule induction machine learning system such as Ripper can perform many times faster than humans.

Regarding which positive/negative example ratio to use in the training data based on the results presented in this paper, the answer is that it depends on how the output is to be used. In a keyword extraction scenario, high precision is preferred over high recall, which means that a low positive/negative ratio would be recommended. In a scenario where the term candidates will be post-processed by domain experts, a high recall is more important that high precision. In that case a balanced ratio such as 50/50 is recommended as this ratio provides a high precision together with a very high recall (as noted in the Results section, 0 false negatives are the result of a "everything-is-a-term system").

class	corr	err	condition
yes	832	273	IF freq ≥ 2.7903e-05 pos1 = n .
yes	403	213	IF logprob ≥ 9.66755 pos1 = n .
yes	387	376	IF pos1 = v .
yes	382	482	IF msd1 = sg-nom .
yes	96	75	IF msd1 = pl-nom .
yes	82	97	IF pos1 = a .
yes	25	11	IF pos1 = n msd1 = sg-gen .
yes	16	5	IF pos1 = n logprob ≥ 9.13763 .
yes	37	51	IF pos1 = ad .
no	812	135	IF .

Table 8: Rules learned in experiment A42B 1-50

class	corr	err	condition
yes	11	4	IF func1 = attr pos1 = <cmp> .
yes	22	15	IF pos1 = a logprob ≥ 12.276 .
yes	5	0	IF func1 = attr pos1 = a logprob ≤ 10.8129 logprob ≥ 10.751 .
yes	11	2	IF logprob ≥ 11.8158 pos2 = adv pos1 = v .
yes	4	1	IF func1 = attr pos1 = <ord> .
yes	3	0	IF msd1 = a-nom .
no	509	3	IF .

Table 9: Rules learned in experiment A61G 2-10

7. Future research

The machine learning approach has been applied to monolingual term extraction in this paper. We have however started working on applying the same approach to bilingual term extraction. Current bilingual term extraction methods, such as presented in (Morin et al., 2007) and (Fan et al., 2009), often rely on monolingual term extraction independently performed on source and target language followed by a term alignment phase between terms in the source and target language (extract-align). There also exist parallel extraction methods that extract terms from aligned texts such as (Merkel and Foo, 2007; Lefever et al., 2009; Foo and Merkel, 2010) (align-extract). A machine learning ap-

	tot	train	train pos	train neg	train pos/neg ratio	test	test pos	test neg	test pos/neg ratio
1-grams	12844	11560	2395	9165	0.261320	1284	485	799	0.607009
2-grams	40221	36199	11	36188	0.000304	4022	9	4013	0.002243

Table 4: A42B experiment data overview

	tot	train	train pos	train neg	train pos/neg ratio	test	test pos	test neg	test pos/neg ratio
1-grams	39243	35319	6239	29080	0.214546	3924	1233	2691	0.458194
2-grams	152853	137568	59	137509	0.000429	15285	30	15255	0.001967

Table 5: A42B experiment data overview

proach presents the opportunity to use a single framework monolingual, bilingual and even multilingual term extraction.

Our current approach uses an n-gram length sensitive feature representation, i.e. a bigram has twice as many linguistic features as a unigram. As a result, for extraction of monolingual terms of length 1 to 3, the machine learning phase has to be split into three separate learning sessions.

Applying this method of representation to bilingual term extraction for terms of length 1 to 3 would require the machine learning phase to be split into nine different sessions (covering 1-1 alignments, 1-2, 1-3 ... 3-3 alignments). Preliminary analysis of bilingually aligned patent texts (English-Swedish) however, indicate that the distribution of aligned phrases according to such a division is very skewed. One way of solving this problem is the change the feature representation to use a single combined feature for all tokens in a phrase rather than separate features for each token in a phrase. This way, the data need not be divided into smaller groups.

8. Conclusion

In this paper, we have shown that machine learning can be used to produce rules which can be used to extract term candidates from a corpus, or more specifically classify n-grams as potential term candidates or not. We have also shown that by using different positive/negative example ratios in the training data, it is possible to govern the kind of rules that are produced. Different kinds of rules may be of interest for different scenarios, so the possibility of dynamically adapting the ruleset to a scenario is promising. Also, we have presented some preliminary results on the use of machine learning for bilingual ATR that indicate that some changes need to be made to to the feature representation scheme to be able to try the approach bilingually.

9. References

Lars Ahrenberg. 2007. Lines 1.0 annotation: Format, contents and guidelines, March.

Sophia Ananiadou. 1994. A methodology for automatic term recognition. In *Proceedings of the 15th conference on Computational linguistics*, pages 1034–1038, Morristown, NJ, USA. Association for Computational Linguistics.

Didier Bourigault. 1992. Surface grammatical analysis for the extraction of terminological noun phrases. In *Proceedings of the 14th conference on Computational linguistics*, pages 977–981, Morristown, NJ, USA. Association for Computational Linguistics.

Jonathan D. Cohen. 1995a. Highlights: language- and domain-independent automatic indexing terms for abstracting. *J. Am. Soc. Inf. Sci.*, 46(3):162–174.

William W. Cohen. 1995b. Fast effective rule induction. In *In Proceedings of the Twelfth International Conference on Machine Learning*, pages 115–123. Morgan Kaufmann.

Béatrice Daille. 1994. Study and implementation of combined techniques for automatic extraction of terminology. In J. Klavans and P. Resnik, editors, *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, pages 49–66. The MIT Press, Cambridge, Massachusetts.

Xiaorong Fan, Nobuyuki Shimizu, and Hiroshi Nakagawa. 2009. Automatic extraction of bilingual terms from a chinese-japanese parallel corpus. In *IUCS '09: Proceedings of the 3rd International Universal Communication Symposium*, pages 41–45, New York, NY, USA. ACM.

Jody Foo and Magnus Merkel. 2010. Computer aided term bank creation and standardization: Building standardized term banks through automated term extraction and advanced editing tools. In Marcel Thelen and Frieda Steurs, editors, *Terminology in Everyday Life*, pages 163–180. John Benjamins Publishing Company.

Eibe Frank, Gordon W. Paynter, Ian H. Witten, Carl Gutwin, and Craig G. Nevill-Manning. 1999. Domain-specific keyphrase extraction. In *IJCAI'99: Proceedings of the 16th international joint conference on Artificial intelligence*, pages 668–673, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Katerina T. Frantzi, Sophia Ananiadou, and Jun ichi Tsujii. 1998. The c-value/nc-value method of automatic recognition for multi-word terms. In *ECDL*, pages 585–604.

Anette Hulth. 2003. Improved automatic keyword extraction given more linguistic knowledge. pages 1–8, January.

Anette Hulth. 2004. *Combining Machine Learning and Natural Language Processing for Automatic Keyword Extraction*. Ph.D. thesis, Stockholm University, Department of Computer and Systems Sciences (together with KTH).

Els Lefever, Lieve Macken, and Veronique Hoste. 2009. Language-independent bilingual terminology extraction

experiment	training data	test data	true pos	true neg	false pos	false neg	recall (terms)	precision (terms)
1-10	10183	1284	336	675	124	149	69.28%	73.04%
1-30	7983	1284	424	511	288	61	87.42%	59.55%
1-50	4790	1284	471	465	334	14	97.11%	58.51%
1-80	2993	1284	485	460	339	0	100.0%	58.86%
1 90	2661	1284	485	0	799	0	100.0%	37.77%
2-10	110	4022	5	3855	158	4	55.56%	3.07%
2-30	36	4022	9	3624	389	0	100.0%	2.26%
2-50	22	4022	9	3624	389	0	100.0%	2.26%
2-80	13	4022	9	3126	887	0	100.0%	1.00%
2-90	12	4022	9	0	4013	0	100.0%	0.22%

Table 6: A42B experiment results

experiment	training data	test data	true pos	true neg	false pos	false neg	recall (terms)	precision (terms)
1-10	32311	3924	682	2252	439	551	55.31%	60.84%
1-30	20796	3924	1069	1600	1091	164	86.70%	49.49%
1-50	12478	3924	1233	0	2691	0	100.0%	31.42%
1-80	7798	3924	1233	0	2691	0	100.0%	31.42%
1-90	6932	3924	1228	941	1750	5	99.59%	41.24%
2-10	590	15285	26	14692	563	4	86.67%	4.41%
2-30	196	15285	30	14360	895	0	100.0%	3.24%
2-50	118	15285	30	13586	1669	0	100.0%	1.77%
2-80	73	15285	30	7928	7327	0	100.0%	0.41%
2-90	65	15285	30	0	15255	0	100.0%	0.20%

Table 7: A61G experiment results

from a multilingual parallel corpus. In *EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 496–504, Morristown, NJ, USA. Association for Computational Linguistics.

Magnus Merkel and Jody Foo. 2007. Terminology extraction and term ranking for standardizing term banks. In Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek, and Mare Koit, editors, *Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007*, Tartu, Estonia. University of Tartu.

Magnus Merkel, Jody Foo, Mikael Andersson, Lars Edholm, Mikaela Gidlund, and Sanna Åsberg. 2009. Automatic extraction and manual validation of hierarchical patent terminology. October.

Emmanuel Morin, Béatrice Daille, Koichi Takeuchi, and Kyo Kageura. 2007. Bilingual terminology mining - using brain, not brawn comparable corpora. In *ACL*.

The parole corpus, at the swedish language bank, university of gothenburg. http://spraakbanken.gu.se/parole/.

J. Ross Quinlan. 1993. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Gerard Salton and Michael J. McGill. 1986. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA.

Peter D. Turney. 2000. Learning algorithms for keyphrase extraction. *Information Retrieval*, 2(4):303–336, May.

Ziqi Zhang, Jose Iria, Christopher Brewster, and Fabio Ciravegna. 2008. A comparative evaluation of term recognition algorithms. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.