Workshop on Creating Cross-language Resources for Disconnected Languages and Styles

Workshop Programme

Sunday, May 27, 2012

14:10 – 14:20 Workshop Presentation

14:20 – 15:00 Session 1: Multilingual Database Generation

Anna Vacalopoulou, Voula Giouli, Eleni Efthimiou and Maria Giagkou, *Bridging the gap between disconnected languages: the eMiLang multi-lingual database*

Sadaf Abdul-Rauf, Mark Fishel, Patrik Lambert, Sandra Noubours and Rico Sennrich, *Extrinsic Evaluation of Sentence Alignment Systems*

15:00 – 16:00 Session 2: Cross-language Resource Derivation

Carolina Scarton and Sandra Aluísio, *Towards a cross-linguistic VerbNet-style lexicon to Brazilian Portuguese*

Parth Gupta, Khushboo Singhal and Paolo Rosso, *Multiword Named Entities Extraction from Cross-Language Text Re-use*

Carlos Rodríguez-Penagos, Jens Grivolla and Joan Codina-Filbá, *Projecting Opinion Mining resources across languages and genres*

16:00 – 16:30 Coffee break

16:30 – 17:30 Session 3: European Projects for Cross-language Resources

Arda Tezcan, Joeri Van de Walle and Heidi Depraetere, *Bologna Translation Service: Constructing Language Resources in the Educational Domain*

Frédérique Segond, Eduard Barbu, Igor Barsanti, Bogomil Kovachev, Nikolaos Lagos, Marco Trevisan and Ed Vald, *From scarcity to bounty: how Galateas can turn your scarce short queries into gold*

Tatiana Gornostay, Anita Gojun, Marion Weller, Ulrich Heid, Emmanuel Morin, Beatrice Daille, Helena Blancafort, Serge Sharoff and Claude Méchoulam, *Terminology Extraction, Translation Tools and Comparable Corpora: TTC concept, midterm progress and achieved results*

17:30 – 18:15 Panel Discussion

18:15 End of Workshop

Workshop Organizing Committee

Patrik Lambert	University of Le Mans, France
Marta R. Costa-jussà	Barcelona Media Innovation Center, Spain
Rafael E. Banchs	Institute for Infocomm Research, Singapore

Workshop Programme Committee

Iñaki Alegria	University of the Basque Country, Spain
Marianna Apidianaki	LIMSI-CNRS, Orsay, France
Victoria Arranz	ELDA, Paris, France
Jordi Atserias	Yahoo! Research, Barcelona, Spain
Gareth Jones	Dublin City University, Ireland
Min-Yen Kan	National University of Singapore
Philipp Koehn	University of Edinburgh, UK
Udo Kruschwitz	University of Essex, UK
Yanjun Ma	Baidu Inc. Beijing, China
Sara Morrissey	Dublin City University, Ireland
Maja Popovic	DFKI, Berlin, Germany
Paolo Rosso	Universidad Politécnica de Valencia, Spain
Marta Recasens	Stanford University, USA
Wade Shen	Massachusetts Institute of Technology, Cambridge, USA

Table of contents

Author Index

Introduction

Linguistic resources have become incredibly valuable as current computational power and data storage capacity have allowed for implementing data-driven and statistical approaches for almost any Natural Language Processing application. Empirical evidence has demonstrated, in a large number of cases and applications, how the availability of appropriate datasets can boost the performance of processing methods and analysis techniques. In this scenario, the availability of data is playing a fundamental role in a new generation of Natural Language Processing applications and technologies.

Nevertheless, there are specific applications and scenarios for which linguistic resources still continue to be scarce. Both, the diversity of languages and the emergence of new communication media and stylistic trends, are responsible for the scarcity of resources in the case of some specific tasks and applications.

In this sense, CREDISLAS aims at studying methods, developing strategies and sharing experiences on creating resources for reducing the linguistic gaps for those specific languages and applications exhibiting resource scarcity problems. More specifically, we focus our attention in three important problems:

- Minority Languages, for which scarcity of resources is a consequence of the minority nature of the language itself. In this case, attention is focused on the development of both monolingual and cross-lingual resources. Some examples in this category include: Basque, Pashto and Haitian Creole, just to mention a few.

- Disconnected Languages, for which a large amount of monolingual resources are available, but due to cultural, historical and/or geographical reasons cross-language resources are actually scarce. Some examples in this category include language pairs such as Chinese and Spanish, Russian and Portuguese, and Arabic and Japanese, just to mention a few.

- New Language Styles, which represent different communication forms or emerging stylistic trends in languages for which the available resources are practically useless. This case includes the typical examples of tweets and chat speak communications, as well as other informal form of communications, which have been recently propelled by the growing phenomenon of the Web2.0.

We hope the CREDISLAS initiative to nourish future research as well as resource development for several useful Natural Language Processing applications and technologies, which should contribute towards a richer heritage of language diversity and availability of linguistics resources for the Natural Language Processing scientific community.

With best regards,
The CREDISLAS organizing team

Patrik Lambert, University of Le Mans
Marta R. Costa-jussà, Barcelona Media Innovation Centre
Rafael E. Banchs, Institute for Infocomm Research

Bridging the gap between disconnected languages: the eMiLang multi-lingual database

Anna Vacalopoulou, Voula Giouli, Eleni Efthimiou, Maria Giagkou
ILSP-Institute for Language and Speech Processing/Athena RC
Artemidos 6 & Epidavrou, Maroussi, Athens, Greece
{avacalop,voula,eleni_e,mgiagkou}@ilsp.gr

Abstract

We present a multi-lingual Lexical Resource (LR) developed in the context of a lexicographic project that involves the development of user-oriented dictionaries for immigrants in Greece. The LR caters to languages that as of yet remain disconnected, and also encompasses a variety of styles that are relevant to communicative situations that the target group is most likely to cope with. We are currently in the process of testing the feasibility to exploit this cross-language and cross-style LR for the automatic acquisition of further large-scale LRs (i.e., comparable corpora), the ultimate goal being to reduce the linguistic gap between the specific disconnected languages and styles.

Keywords: cross-language resources, disconnected languages, disconnected styles, dictionaries, corpus-based lexicography

1. Introduction

Developing Language Resources (LRs) is a laborious task usually hampered by a lack of available data in the appropriate specialised domains. Furthermore, construction and collection of cross-language resources for applications such as Machine Translation (MT) or Cross-Language Information Retrieval (CLIR) is even more problematic, especially when language pairs not involving English are concerned. State-of-the-art data-driven methods that are currently adopted in multi- or cross-lingual applications directly depend on the availability of very large quantities of parallel LRs. In effect, the accuracy of such data-driven applications varies significantly from being quite good for well-represented languages and language pairs (e.g. English, French, Chinese, Arabic, German, etc.) to being far below acceptable for under-resourced *languages* (or language pairs) and *domains*. Finally, *portability* of existing tools to new languages and domains depends on the availability of appropriate data.

This paper describes a multilingual lexical database that connects language pairs that as of yet remain unconnected, and also the approach we have adopted to the acquisition of multilingual comparable corpora from the web by exploiting and enhancing this resource.

The reported work was carried out in the framework of the national Greek project *eMiLang* (GSRT), aiming at supporting linguistic adaptation of immigrant populations in Greece.

2. The *eMiLang* Dictionaries

eMiLang is a project in progress, aiming to develop a *digital infrastructure* for the support of adult immigrants in Greece. The ultimate goal of *eMiLang* is to assist both immigrants and policy makers in their joint efforts for smooth integration of the target groups to the Greek society. The intended infrastructure encompasses two interrelated pillars: (a) the development of specialized multilingual parallel corpora in the form of informative material and bilingual dictionaries (extracted, in part, from these corpora), and (b) the implementation of a multilingual, multimedia web interface, designed so as to integrate the aforementioned digital content in its entirety. This interface will also offer advanced search mechanisms and information retrieval capabilities. Finally, a news aggregator will be integrated into the system, offering digital information services to the users. This paper will describe the creation of the *eMiLang* dictionaries and their experimental use and reusability in reducing the linguistic gap between disconnected languages and styles.

2.1. The eMiLang Dictionaries as Cross-Language Resources

The *eMiLang* dictionaries (Vacalopoulou et al., 2011) cover the most common range of foreign languages used and/or understood currently by the majority of the immigrant community in Greece.[1] Thus, nine bilingual dictionaries are created, namely: *Greek-Albanian (EL-AL), Greek-Arabic (EL-AR), Greek-Bulgarian (EL-BG), Greek-Chinese (EL-CH), Greek-English (EL-EN), Greek-Polish (EL-PL), Greek-Romanian (EL-RO), Greek-Russian (EL-RU),* and *Greek-Serbian (EL-SR)*. Each bilingual dictionary comprises approximately 15,000 entries which cover mainly the basic vocabulary of Greek. Although a formal complete list of basic Greek vocabulary is still missing from the literature, in the current implementation, the basic vocabulary is conceived as one which comprises not only the most frequent items but also less frequent words and phrases that are relative to everyday life.

2.2. The eMiLang Dictionaries as Cross-Style Resources

Apart from the *basic vocabulary*, another substantial category of lemmas is the one often occurring in official, administrative or other documents which the target group

[1] The selection of languages was based in an extensive comparison of numerical data provided both by Eurostat (http://epp.eurostat.ec.europa.eu/) and by the Hellenic Statistical Authority (http://www.statistics.gr/).

is likely to encounter during their stay in Greece, as for example when applying for a residence permit. To this end, selected *technical vocabulary*, that is, terms pertaining to domains/subject fields that are of utmost interest to the target group have been included as well.

Because of the fact that the target group is generally expected to lack basic encyclopaedic information about Greece, the dictionaries also contain proper nouns. These include the names of: (a) *geographical entities* (i.e. cities, islands, regions etc.), (b) *official bodies* (i.e. ministries and other official organisations), and (c) *geopolitical entities* (*Ηνωμένα Έθνη = United Nations*). Both official bodies and geopolitical entities are quite often expressed by acronyms which are also retained in the lemma list.

The process of dictionary compilation has been corpus-based; this refers to headword selection, sense disambiguation and extraction of collocations and usage examples. Dictionary entries were semi-automatically selected from a variety of sources, including (a) a large (POS-tagged and lemmatized) reference corpus of the Greek language, namely the Hellenic National Corpus (http://hnc.ilsp.gr/), (b) a specialized Greek corpus specially collected within the framework of the current project, that adheres to pre-defined domains (administrative, culture, education, health, travel, and welfare), and (c) already existing dictionaries and glossaries, customized to better suit the user needs (communicative situations and relevant vocabulary, etc.). Such resources have been previously developed by ILSP for the purpose of other projects and they are either published [2] of non-published works. As a result, a proportion of the entries is part of what can be conceived as the *basic vocabulary* of Greek. This does not only mean the most frequent items attested in the HNC, but also less frequent words and phrases that are relative to everyday life, and which are used to populate the domains described above (such as *μαξιλαροθήκη = pillowcase* or *πάνα = nappy*).

Furthermore, the dictionaries follow the closed vocabulary concept, thus including every word in the examples as an entry itself for easy reference. This has led to adding a considerable amount of *entries ad hoc* and keeping a better balance, in terms of content, between everyday vocabulary and the administrative jargon of the public service.

3. Standards for resource creation

It is evident from the above that the intended resource will not only bring together disconnected languages but also very disconnected styles. It has been decided that a certain set of rules were to be followed, in order to meet this double challenge. First, as the dictionaries are mainly targeted towards starter learners of Greek who are in need of speedy learning, it has been decided that only basic meanings would be included in it. Meanings are implicitly presented through one or more examples of usage, which bear the informative load. Examples of usage are thus a core element of the dictionary. Furthermore, dictionary examples have been carefully selected so as to reflect not only the different meanings but also the most basic forms of usage, grammar and/or collocation. Thus, for instance, the active and passive of verbs are presented separately when voice differentiates meaning as well; the same stands for verbs used with different prepositions etc.

As the emphasis of these dictionaries has been to include as much information as possible but in the most user-friendly way possible, examples have been selected so as to be as interesting as possible to the target group. To this end, a combination of different corpora (mentioned earlier) has been used. Thus, a large part of the examples for the basic vocabulary was extracted from the Hellenic National Corpus, although usually shortened and/or simplified to suit the target group level.

In terms of length, examples are short and contain no excess information. They usually consist of one simple sentence, although some dialogue is included to exemplify everyday phrases, such as greetings or asking for information. Apart from accelerating the learning process, the brevity criterion also simplifies the ambitious work of translating everything into 9 languages.

As it is customary in most multilingual dictionaries, examples also play the role of describing each meaning, due to lack of definition. This has placed additional difficulty in selecting the right example for each meaning. For instance, an example of the verb *αγωνίζομαι = struggle* would be *Αγωνίστηκε πολύ, για να καταφέρει αυτό που ήθελε = She struggled a lot to get what she wanted*.

Last but not least, taking into account the great variety of backgrounds from which the target group of this dictionary comes, extra care has been taken towards political correctness. All examples are free of any social, political, racial, national, and religious or gender bias.

4. Bootstrapping Language Resources

As it has been pointed out, the LR described above has been developed in a specific context and for particular purposes. The languages that were handled within this project are to a great extent disconnected. However, the development of these bilingual dictionaries may be considered as the primary step towards developing further resources (comparable corpora, bilingual lexica) semi-automatically from sparse data. To this end, we argue that this resource can be repurposed in view of bootstrapping the acquisition of mono- and cross-lingual corpora that might be useful in a range of NLP applications from Machine Translation to Cross-Lingual Information Extraction, etc., and for language pairs and domains that to-date remain disconnected.

4.1 Comparable corpora: a means to connect disconnected languages

The problem of the limited availability of linguistic

[2] For example: Electronic Greek-Turkish dictionary for young learners, Athens 2004 XENION Lexicon, Athens 2005.

resources is especially relevant for language pairs that involve either less-resourced or disconnected languages. A number of surveys[3] on existing corpora have revealed the availability of parallel corpora, yet, in most cases, English has been used as pivot, and other languages (including the ones mentioned here) remain disconnected. Moreover, parallel textual resources comprise mainly of bilingual texts in the "resource-affluent" languages, i.e. English, French, German, Arabic and Chinese (Mihalcea et al, 2005). Additionally, Gavriilidou et al. (2006) have identified a number of drawbacks that challenge the identification of parallel texts for less-resource languages the most obvious being the real status of the web, which is attested to be multilingual but not parallel: parallel texts in multiple languages are extremely rare, especially for the less widely available ones, given that organizations and multilingual portals usually provide their content mainly in the most dominant languages. A similar position is hinted at in (Resnik & Smith, 2003) and (Mihalcea & Simard, 2005). Additionally, lack of "true parallelness" of the web in the sense that seemingly parallel texts are usually proved to be only partially parallel, while, quite often, "translations" prove to be summaries or paraphrases of the original text also hampers the acquisition of parallel data.

On the other hand, although large-scale multilingual corpora, as for example, the *Europarl* parallel corpus (Kehn, 2005) and the aligned multilingual parallel corpus *JRC-ACQUIS*, contain many language combinations, yet they are domain-specific. Adaptation to new domains requires extra efforts.

In recent years, comparable corpora have been considered as a means to accommodate the scarcity of parallel ones. In this sense, comparable corpora can be seen as a means to bridge disconnected languages. A comparable corpus in contrast to a parallel one, is generally defined as a collection of documents that are gathered according to a set of criteria along the axes of content and time, i.e., the corpus must contain the same proportion of texts of the same *genre* in the same *domains* in the same *period* in two or more languages (McEnery and Xiao, 2005). It has been proven (Munteanu and Marcu, 2005), (Munteanu, 2006), (Maia and Matos, 2008), (Hewavitharana and Vogel, 2008), (Goutte et al., 2009) that comparable corpora can compensate for the shortage of parallel ones since training data that has a significant impact on the performance of SMT (Statistical Machine Translation) systems can be extracted from them.

Other uses of comparable corpora can be seen in non-machine translation (Kubler, 2008), and even in language learning (Bacelar do Nascimento et al., 2008), etc. Moreover, large collections of raw data can be automatically annotated and used to produce, by means of induction tools, a second order or synthesized derivatives: rich lexica (with morphological, syntactic and lexico-semantic information) and massive bilingual dictionaries (word and multiword based) and transfer grammars. Finally, bilingual lexicon extraction from non-aligned comparable corpora, phrasal translation as well as evaluations on Cross-Language Information Retrieval may be seen as possible use cases of comparable corpora in view of connecting disconnected languages in a number of settings.

4.2 Using eMiLang data to collect in-domain comparable corpora

Within the current research, the feasibility to bootstrap comparable corpora from the web sources by exploiting eMiLang dictionaries has been tested. In this section, we will elaborate on the methodology employed to conduct a pilot research.

Corpus collection from web sources has been attempted using a crawler (Mastropavlos et al, 2011), i.e. an engine that was developed at the Institute for Language and Speech Processing, which starts from a few seed URLs and "travels" on the Web to find web pages in the targeted languages that are relevant to specific domains. The crawler attempts to fetch monolingual documents from these web pages by making use of topic definitions, i.e., weighted lists of terms that are relevant to the specific domain. After crawling, text normalization, cleaning and de-duplication are the main subtasks involved in the automatic construction of the corpora. The text normalization phase involves detection of the format and text encoding of the downloaded web pages and conversion of these pages into plain text and text encoding (UTF-8).In the remainder of the document, we will describe the procedure followed for creating the topic definitions and we will discuss initial results.

The required input for the crawler consists of a topic definition and a list of seed URLs in the languages involved in a given task. The creation of these language and domain-specific resources is an off-line task that requires manual effort. Being a critical issue to the acquisition of bilingual comparable corpora, the construction of topic definitions that are similar across languages, exploited the bilingual eMiLang dictionaries.

More precisely, EL single- and multi-words included in the dictionaries and pertaining to pre-specified domains, namely, *Administrative*, *Finance*, *Foods/Nutrition*, *Health/Fitness*, *Law* and *Transport* were initially extracted from the multi-lingual database on the basis of the domain labels that they were assigned (cf. above). The so-extracted lists were further processed manually in view of creating the EL domain-specific topic definitions. This processing was kept to a minimum and was meant to remove duplicate entries (featuring different Part-of-Speech usages) and to omit terms representing concepts that are culture-specific (and therefore probably not lexicalized in the target languages). On top of that, conformance with the pre-defined format was ensured, since we adopted a widely-accepted strategy (Ardo and Golub, 2007), (Dorado, 2008), i.e. to use triplets (<term, relevance weight, topic-class >) as the basic entities of the

[3] These include surveys conducted in the framework of the ENABLER project (Calzolari et al., 2004), surveys carried out by organizations involved in LRs activities such as ELRA (http://www.elra.info/).

topic definition.

Consequently, weights were assigned to the EL terms semi-automatically on the basis of their frequency of occurrence in the relevant sub-corpora obtained out of a monolingual corpus of contemporary Greek, namely, the Hellenic National Corpus (Hatzigeorgiu et al., 2000). Weights are signed integers that indicate the relevance of the term with respect to the domain. Higher values indicate more relevant terms. The construction of *topic definitions* for each domain separately was finalized by adding an appropriate domain or sub-domain value, as for example: *administrative,* or *administrative-politics* as appropriate.

The so-constructed EL *topic definitions* along with seed URLs that were appropriately selected were then fed to the crawler and the monolingual in-domain EL corpora were thus obtained for the aforementioned domains.

To cater for the collection of similar in-domain corpora in the target languages that would be comparable to the already acquired EL ones, the construction of *topic definitions* in the target languages was in order. Since no translations are provided in the dictionaries at the entry level, the EL examples relevant to each entry/sense along with their translations in each one of the target languages were exploited. Additionally, inflectional forms of entries' translational equivalents were supplied by the translators during translation and further exploited. Finally, weighs and domain/subdomain values were retained from the EL *topic definition*, whereas seed URLs were also provided by native informants.

The lists of seed URLs were collected from various web directories updated by human editors. The seed URLs employed in this setting were selected from the lists available in the Open Directory Project (ODP)[4]. It should also be noted that we tried to select comparable sources across languages: administrative bodies, organizations, financial portals and newspapers, etc.

Finally, documents delivered are retained in XML format, which is compatible to the Corpus Encoding Standard (CES)[5]. These files contain metadata elements (title, distributor, source url, text language, domain information, and topics identified in the text). Moreover, text is segmented in paragraphs.

4.3 Results - discussion

We conducted a suite of pilot experiments on the EL and EN data in two domains, namely, *administrative* and *finance* in order to check the feasibility of the endeavour. The initial results reported in Table 1 below reflect the crawler's running for about 1 hour. 179 EL and 180 EN terms pertaining to the Administrative domain yielded 1640 and 1890 files in Greek and English respectively. Similarly, 99 EL and 76 EN terms in the Finance domain returned 1823 and 2021 files in Greek and English financial documents. The seed URLs lists contained 6 websites per language and domain.

A sub-part of the so-collected EL and EN monolingual corpora were then hand-validated with respect to domain-specificity. Manual validation consisted in deciding whether documents retrieved using the methodology described above were accurately classified as in-domain or not.

Domain	**EL**		**EN**	
	files	terms	files	terms
Admin	1640	179	1890	180
Finance	1823	99	2021	76

Table 1: Crawling data

So far, c. 50% of the retrieved files that were identified as pertaining to either the selected domains was checked manually and initial results seemed to be encouraging (0.68 accuracy).

There are, however, open issues with respect to genre or text type identification. A closer inspection over the problematic cases showed that in most cases, false positives contained single- and multi-word *entries* that were included in the topic definition whose *termhood* is somehow disputable. This was due to the fact that general-language words are also classified invariably as pertaining to a domain. For example *δεκάευρο (= ten euro note)* is classified as *finance.* As a result, documents including this term were not applicable for building a strictly financial corpus. To this end, we believe that modification or fine-tuning of the relative weighs along with strict selection of seed URLs is in order so as to effectively collect comparable corpora that take text type and genre into account. Further experimentation has been planned so as to check the feasibility of this assumption as well.

5. Conclusions – Future research

We have presented a multi-lingual lexical database that was initially developed manually in the framework of a lexicographic project. The lexical database covers language pairs that as yet seemed to be unconnected, especially for specialised domains. The resource is being used to bootstrap the automatic acquisition of comparable corpora in the languages involved and in pre-defined domains from web sources. For the time being, initial experiments in specialised domains in *EL* and *EN* have proven quite promising. Future work involves the acquisition of corpora in the remaining languages and domains. Moreover, multiple iterations of the procedure and enrichment of topic definitions with new lexical entries extracted from the acquired corpora will be also attempted. As it has already been mentioned above, further experimentation has been planned so as to better exploit the cross-style feature of the resource.

The obvious next step will be the exploitation of these comparable corpora for the extraction of bilingual LRs (terminological, phrasal, etc) that would be of interest to a

[4] Open Directory Project: http://www.dmoz.org/

[5] XML Corpus Encoding Standard Document XCES 1.0.4 Last Modified 19 February 2009

number of NLP applications.

6. Acknowledgements

The research within the project "*eMiLang: Improved Informative Digital Services Targeted on Immigrants*" leading to these results has received funding from the Greek state, under the framework of *O. P. Digital Convergence (2007-2013).*

7. References

Ardo, A., and Golub, K. (2007). Documentation for the Combine (focused) crawling system, http://combine.it.lth.se/documentation/DocMain/

Atkins, S.B.T. (1998). Using Dictionaries: Studies of Dictionary use by Language Learners and Translators.Tübingen: Max Niemeyer Verlag.

Baldwin - Edwards, M. (2008). Immigrants in Greece: Characteristics and Issues of regional distribution. MMO Working Paper No. 10, Jan. 2008

Baldwin - Edwards, M. (2004). Statistical Data on Immigrants in Greece, Athens: Mediterranean Migration Observatory and IMEPO.

Bacelar do Nascimento M., Estrela A., Mendes A., Pereira L. (2008). On the use of comparable corpora of African varieties of Portuguese for linguistic description and teaching/learning applications. In *Proceedings of the Workshop on Building and Using Comparable Corpora, Language Resources and Evaluation Conference, Marrakech, Morocco*.

Bekavac B., Osenova P., Simov, K., Tadić, M. (2004) Making Monolingual Corpora Comparable: a Case Study of Bulgarian and Croatian. In *Proceedings of the 4th Language Resources and Evaluation Conference: LREC04*, Lisbon, Portugal.

Calzolari, N., Choukri, K., Gavrilidou, M., Maegaard, B., Baroni, P., Fersoe, H., Lenci, A., Mapelli, V., Monachini, M., Piperidis, S. (2004). ENABLER Thematic Network of National Projects: Technical, Strategic and Political Issues of LRs. In *Proceedings of the 4th Language Resources and Evaluation Conference: LREC04*, Lisbon, Portugal.

Daille, B., and Morin, E. (2005). French-English Terminology Extraction from Comparable Corpora. In *R. Dale et al. (Eds.): IJCNLP 2005, LNAI 3651, pp. 707–718, 2005*.

De Schryver, G.M., and Prinsloo, D.J. (2000). Dictionary-Making Process with 'Simultaneous Feedback' from the Target Users to the Compilers. *Lexikos 10*: 1–31.

Dorado, I. G. (2008). Focused Crawling: algorithm survey and new approaches with a manual analysis. Master thesis.

Gavrilidou, M., Labropoulou, P., Piperidis, S., Giouli, V., Calzolari, N., Monachini, M., Soria, C., Choukri, K., 2006. Language Resources Production Models: the Case of the INTERA Multilingual Corpus and Terminology. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Geneva, Italy.

Goutte C., Cancedda N., Dymetman M., Foster G. (eds.) (2009) Learning Machine Translation. The MIT Press. Cambridge, Massachusetts, London, England.

Hatzigeorgiu N., Gavrilidou M., Piperidis S., Carayannis G., Papakostopoulou A., Spiliotopoulou A., Vacalopoulou A., Labropoulou P., Mantzari E., Papageorgiou H., Demiros I. (2000). Design and implementation of the online ILSP Greek Corpus. In *Proceedings of the 2nd Language Resources and Evaluation Conference: LREC00*, Athens, Greece.

Hewavitharana S. and Vogel S. (2008) Enhancing a Statistical Machine Translation System by using an Automatically Extracted Parallel Corpus from Comparable Sources. In *Proceedings of the Workshop on Building and Using Comparable Corpora, Language Resources and Evaluation Conference, MarrakechMorocco, 2008*.

Koehn, Philipp. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation, , MT Summit 2005

Kübler N. (2008) A comparable Learner Translator Corpus: creation and use. In *Proceedings of the Workshop on Building and Using Comparable Corpora, Language Resources and Evaluation Conference, MarrakechMorocco, 2008*.

Maia B. and Matos S. (2008) Corpógrafo V.4 – Tools for Researchers and Teachers Using Comparable Corpora. In *Proceedings of the Workshop on Building and Using Comparable Corpora, Language Resources and Evaluation Conference, MarrakechMorocco, 2008*.

Mastropavlos, N., Papavassiliou, V. (2011). Automatic Acquisition of Bilingual Language Resources. In *Proceedings of the 10th International Conference of Greek Linguistics, Komotini, Greece.*

McEnery A.M., Xiao R.Z. (2007). Parallel and comparable corpora: What are they up to? In *Incorporating Corpora: Translation and the Linguist. Translating Europe*. Multilingual Matters, Clevedon.

Mihalcea, R., Simard, M. (2005). Parallel Texts. *Journal of Natural Language Engineering*, 11(3), pp. 239-246.

Munteanu D. (2006). Exploiting Comparable Corpora (for automatic creation of parallel corpora). Online presentation.http://content.digitalwell.washington.edu/msr/external_release_talks_12_05_2005/14008/lecture.htm

Munteanu D., Marcu D. (2005) Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. Computational Linguistics, 31(4), pp. 477-504.

Vacalopoulou, A., Giouli, V., Giagkou, M., and Efthimiou, E. 2011. Online Dictionaries for immigrants in Greece: Overcoming the Communication Barriers. In *Proceedings of the 2nd Conference "Electronic Lexicography in the 21st century: new Applications for New users"* (eLEX2011), Bled, Slovenia.

Varantola, K. (2002). Use and Usability of Dictionaries: Common Sense and Context Sensibility?. In Marie-Helene Correard (editor). Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins. Grenoble, France: EURALEX, 2002.

Extrinsic Evaluation of Sentence Alignment Systems

Sadaf Abdul-Rauf[1], Mark Fishel[2], Patrik Lambert[1], Sandra Noubours[3], Rico Sennrich[2]

[1]LIUM, University of Le Mans, France
sadaf.abdul-rauf,patrik.lambert@lium.univ-lemans.fr

[2]Institute of Computational Linguistics, University of Zurich, Switzerland
sennrich,fishel@cl.uzh.ch

[3]Fraunhofer FKIE, Wachtberg, Germany
sandra.noubours@fkie.fraunhofer.de

Abstract

Parallel corpora are usually a collection of *documents* which are translations of each other. To be useful in NLP applications such as word alignment or machine translation, they first have to be aligned at the sentence level. This paper is a user study briefly reviewing several sentence aligners and evaluating them based on the performance achieved by the SMT systems trained on their output. We conducted experiments on two language pairs and showed that using a more advanced sentence alignment algorithm may yield gains of 0.5 to 1 BLEU points.

Keywords: sentence alignment, parallel corpora, evaluation

1. Introduction

Parallel corpora[1] constitute an essential cross-language resource whose scarcity for a given language pair and domain restricts the development of data-driven natural language processing (NLP) approaches for that language pair and domain. In this respect, building a parallel corpus helps connecting the considered languages.

After collection, the size of the translated segments forming the parallel corpus are usually of the order of entire documents (e.g. European Parliament sessions or newspaper articles). Learning word correspondences with this kind of examples is an ambiguous task. The ambiguity may be reduced by first decreasing the size of the segments within each pair. This task is called sentence alignment and consists of finding correspondences between segments such as sentences or small paragraphs within a pair of translated documents. The existence of (meta-)textual information such as time stamps (subtitles), speaker information (Europarl (Koehn, 2005)), or paragraphs/chapters/smaller documents provides anchors at which the two sides are certainly aligned. It may thus considerably reduce the complexity of the sentence alignment task. The more fine-grained we can align the text based on textual structure, the easier sentence alignment becomes.

This paper details a user study initiated at the 5th Machine Translation Marathon[2], and whose aim was to evaluate sentence alignment tools on different types of document-aligned parallel corpora and measure its impact on an NLP task, namely Statistical Machine Translation (SMT). The test was conducted on two language pairs. First, on NIST 2008 Urdu–English training data, which contains documents of about 17 sentences in average, with no informative meta- or textual information. Second, on the concatenation of three collections of French–English texts:

- the BAF corpus,[3] composed of very long documents (thousands of lines) with few possible anchors in the text.
- the News Commentary corpus, a corpus of news commentary articles crawled from the web[4], with HTML paragraph mark-up information.
- a corpus crawled from Rapid[5], a site with press releases of the European Union (also containing paragraph mark-up information)

We evaluated five unsupervised sentence alignment tools: the Gale and Church algorithm, Microsoft's Bilingual Sentence Aligner (MBA), Hunalign, Gargantua and Bleualign. In the next section, we describe these sentence alignment tools. Then we present experimental results obtained on the Urdu–English and French–English data. Finally, we draw some conclusions.

2. Sentence Alignment Tools

All five sentence alignment tools that we evaluated use a dynamic programming search to find the best path of sentence pairs through a parallel text. This means that all of them assume that the texts are ordered monotonically and none of the tools is able to extract crossing sentence pairs. For texts with major changes in sentence order between two language version, parallel sentence extraction may be preferrable to searching a global sentence alignment (Fung and Cheung, 2004). All tools also resort to some pruning strategy to restrict the search space.

[1] A parallel corpus is a collection of segment pairs, the two segments within each pair being translation of each other.

[2] http://lium3.univ-lemans.fr/mtmarathon2010/

[3] http://rali.iro.umontreal.ca/Ressources/BAF/

[4] http://www.project-syndicate.org/

[5] http://europa.eu/rapid

While some of the tools support the use of external resources (i.e. bilingual dictionaries in the case of Hunalign, and existing MT systems for Bleualign), all systems learned their respective models from the parallel text itself.

2.1. Gale and Church Algorithm

The Gale and Church (1991; 1993) algorithm is based on character based sentence length correlations, i.e. the algorithm tries to match sentences of similar length and merges sentences, if necessary, based on the number of words in the sentences. The alignment model proposed by Gale and Church (1993) makes use of the fact that longer/shorter sentences in one language tend to be translated into longer/shorter sentences in the other. A probabilistic score is assigned to each proposed sentence pair, based on the sentence length ratio of the two sentences (in characters) and the variance of this ratio. This probabilistic score is then used in the dynamic programming framework to get the maximum likelihood alignment of sentences. Some corpora aligned using this algorithm include the Europarl corpus (Koehn, 2005) and the JRC-Acquis (Steinberger et al., 2006) among others.

2.2. Bilingual Sentence Aligner (MBA)

The Bilingual Sentence Aligner (Moore, 2002) combines a sentence-length-based method with a word-correspondence-based method. While sentence alignment based on sentence-length is relatively fast, lexical methods are generally more accurate but slower. Moore's hybrid approach aims at realising an accurate and computationally efficient sentence alignment model that is not dependent on any additional linguistic resources or knowledge.

The aligner implements a two-stage approach. First the corpus is aligned based on sentence length. The sentence pairs that are assigned the highest probability of alignment are then used as training data for the next stage. In this second stage, a lexical model is trained, which is a modified version of IBM model 1. The final alignment model for the corpus combines the initial alignment model with IBM model 1. These alignments are therefore based on both sentence length and word correspondences and comprise 1-to-1 correspondences with high precision.

2.3. Hunalign

Hunalign (Varga et al., 2005) implements an alignment algorithm based on both sentence length and lexical similarity. It is thus in general similar to Moore's algorithm. The main difference is that Hunalign uses a crude word-by-word dictionary-based replacement instead of IBM model 1. On one hand this results in significant speed gains. More importantly, however, it provides flexible dependence on the dictionary, which can be pre-specified (if one is available) or learned empirically from the data itself.

In case a dictionary is not available, an initial pass is made, based only on sentence length similarity, after which the dictionary is estimated from this initial alignment and a second pass, this time with the dictionary is made.

Although Hunalign is optimised for speed, its memory consumption is its weak spot; in reality it cannot handle parallel corpora larger than 20 thousand sentences – these have to be split into smaller chunks, which results in worse dictionary estimates.

Language	Docs	Max Len.	Ave. Len.	Segm.	Words
Urdu	5282	1003	17.7	93 332	1800 k
English	5282	878	16.9	89 323	2027 k
French	3461	7077	54.2	187 656	4104 k
English	3461	6890	54.1	187 213	3486 k

Table 1: Statistics for the training data set for NIST Urdu–English data and for the French–English data (k stands for thousands).

2.4. Gargantua

Gargantua (Braune and Fraser, 2010) aims to improve on the alignment algorithm by Moore (2002) by replacing the second pass of Moore's algorithm with a two-step clustering approach. As in Moore's algorithm, the first pass is based on sentence-length statistics and used to train an IBM model. The second pass, which uses the lexical model from the first pass, consists of two steps. In a first step, a sequence of 1-to-1 alignments is obtained through dynamic programming. In a second step, these are merged with unaligned sentences to build 1-to-many and many-to-1 alignments.

2.5. Bleualign

Bleualign (Sennrich and Volk, 2010) uses an automatic translation of the source text as an intermediary between the source text and the target text. A first alignment is computed between the translated source text and the target text by measuring surface similarity between all sentence pairs, using a variant of BLEU, then finding a path of 1-to-1 alignments that maximises the total score through dynamic programming. In a second pass, further 1-to-1, many-to-1 and 1-to-many alignments are added through various heuristics, using the alignments of the first pass as anchors.

Bleualign does not build its own translation model for the translation of the source text, but requires an external MT system. In order not to skew the evaluation by using additional resources, we followed Sennrich and Volk (2011) in performing a bootstrapped alignment. As a first step, we aligned the parallel text with the Gale & Church algorithm. Then, we built a SMT system out of this aligned parallel text, and automatically translated the (unaligned) source text. This translation is the basis for the final alignment with Bleualign.

3. Experiments

The aim of the study was to use each sentence aligner to find correspondences at the sentence level in a document-aligned parallel corpus. Then an SMT system was trained from the resulting sentence-aligned parallel corpus, tuned on a development set and used to translate a test set. The sentence aligners were evaluated based on the quality of the translation with respect to automated metrics. The experiment was conducted on two language pairs. The statistics of the document-aligned training data for each language

Set	Language	Segments	Words	Vocabulary	Lmean	Ref.
Dev.	Urdu	923	28.1 k	5.4 k	30.3	1
1st ref.	English	923	24.2 k	5.0 k	26.3	
Test	Urdu	1862	42.3 k	6.5 k	22.7	4
1st ref.	English	1862	38.2 k	6.2 k	20.5	
Dev.	French	2051	55.4 k	9.2 k	27.0	1
1st ref.	English	2051	49.8 k	8.4 k	24.3	
Test	French	2525	72.5 k	11.2 k	28.7	1
1st ref.	English	2525	65.6 k	9.7 k	26.0	

Table 2: Basic statistics for the translation system development and test data sets (k stands for thousands, Lmean refers to the average segment length in number of words, and Ref. to the number of available translation references).

pair are presented in Table 1. These statistics are the number of documents, the maximum document length and the average document length in segments, the total number of segments and the total number of running words in the corpus. The statistics of the development and test data for the SMT systems are presented in Table 2. The statistics shown are the number of segments, the number of words, the vocabulary size (or number of distinct words), the average segment length in number of words and the number of available translation references.

3.1. Urdu–English Task

The Urdu–English data presented in Tables 1 and 2 were provided at NIST 2008 Machine Translation evaluation.[6] The available parallel training and development corpora were only aligned at the document level. We used the training data for the unsupervised sentence alignment. We aligned a part of the development data at the sentence level with the Bleualign tool to build a corpus to tune the SMT systems (Urdu “Dev.” in Table 2). Our test set for extrinsic evaluation was the official NIST 2008 test set (Urdu “Test” in Table 2).

The output of the sentence aligners contains at most the same number of tokens as in the training corpus. For some segments, they indeed fail to find any corresponding segment in the other side of the corpus. Table 3 indicates the coverage in terms of number of tokens achieved by the various aligners tested. The % columns indicate the percentage of tokens in the sentence aligned parallel texts compared to the original amount in the training corpus. Gale and Church, Gargantua and Hunalign achieved a coverage around 95%. Bleualign achieved a slightly lower coverage (close to 90%). The MBA only output less than 45% of the input tokens. This can be explained by two reasons. First, it was used with its default precision threshold, which was particularly selective because the Urdu–English data may be noisy or not strictly parallel. A different threshold could have allowed the tool to achieve a higher coverage. Second, the MBA can only extract 1-to-1 correspondences.

The parallel texts described in Table 3 were used to train phrase-based SMT systems with the Moses toolkit (Koehn et al., 2007). In order to stick to the tight MT Marathon schedule, we used an existing language model, trained with news data and data from the European Parliament and the United Nation proceedings.[7] Thus the target side of the sentence-aligned training corpus may not be included in the language model training data. Table 4 shows the scores of three automated MT metrics, namely BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and TER (Snover et al., 2006), obtained by the SMT system trained on the output of each sentence aligner. The evaluation was case-sensitive. The values shown are the average and standard deviation over 3 MERT runs with different random seeds. The values in bold are possibly the best one taking the error range into account.

	Segments (k)		Tokens (k)			
	Urdu	English	Urdu	%	English	%
Training	93.3	89.3	2027	100.0	1800	100.0
Bleualign	65.6		1821	89.9	1607	89.3
Gale&Church	70.0		1925	95.0	1729	96.1
Gargantua	71.1		1943	95.9	1737	96.5
Hunalign	68.7		1950	96.2	1670	92.8
MBA	40.3		902	44.5	745	41.4

Table 3: Coverage on Urdu–English data

Aligner	BLEU	METEOR	TER
Bleualign	**18.1 ±0.3**	**36.0 ±0.2**	**67.9 ±0.7**
Gale&Church	17.0 ±0.3	**35.6 ±0.7**	70.8 ±1.0
Gargantua	**18.1 ±0.2**	**35.6 ±0.4**	**68.1 ±0.7**
Hunalign	17.1 ±0.4	35.3 ±0.2	**69.5 ±1.4**
MBA	17.2 ±0.2	35.4 ±0.2	70.9 ±0.8

Table 4: SMT results on Urdu–English data.

Bleualign and Gargantua tools achieved the highest rank according to all three metrics. Gale and Church and Hunalign methods ranked first according to only one metric. With the corresponding SMT system trained on half the data, MBA achieved worse scores than the other tools according to all metrics. However, the relative difference was below 5%. Still, on this data set one can achieve a significant performance gain by using one of the best tools versus using one of the most basic ones (about 1 BLEU point, 0.5 Meteor point and more than 1.5 TER point).

[6] http://www.itl.nist.gov/iad/mig/tests/mt/2008/

[7] These data are available at http://www.statmt.org/wmt10/.

3.2. French–English Task

We repeated our study on the French–English data, whose statistics are presented in Tables 1 and 2. The training corpus for sentence alignment was described in Sect. 1. The development (French "Dev." in Table 2) and test data (French "Test" in Table 2) were respectively the test set of the 2008 and 2009 Workshop of Statistical Machine Translation shared tasks (see footnote 7).

Table 5 indicates the coverage achieved by the various aligners tested on the French–English data. With this data set the coverage is higher, and the difference between aligners is lower. In particular, the MBA coverage is only 13% lower than that of the aligner with best coverage.

	Segments (k)		Tokens (k)			
	French	English	French	%	English	%
Training	187.7	187.2	4105	100.0	3487	100.0
Bleualign	140.7		3962	96.5	3392	97.3
Gale&Church	141.6		4022	98.0	3440	98.7
Gargantua	142.4		4005	97.6	3430	98.4
Hunalign	142.4		3996	97.4	3414	97.9
MBA	131.7		3503	85.3	3014	86.4

Table 5: Coverage on French–English data

Table 6 shows the (case-sensitive) scores of automated MT metrics achieved by the SMT systems trained (in the same way as in Sect. 3.1.) on the French–English parallel texts output by the different sentence aligners. On this task the difference between aligners is lower than on the Urdu–English task. This may be explained by the presence in a part of the corpus of HTML mark-up information, such as paragraphs, sub-sections or links, which makes the sentence alignment task easier. By using the best aligner instead of the worst one, one can achieve a gain of 0.4 BLEU point, 0.3 Meteor point and 0.5 TER point. Bleualign and Hunalign ranked first according to all three metrics. Gargantua and MBA ranked first according to one metric, and the Gale and Church method did not rank first at all.

Aligner	BLEU	METEOR	TER
Bleualign	**21.07 ±0.07**	**38.83 ±0.15**	**61.2 ±0.2**
Gale&Church	20.64 ±0.07	38.54 ±0.15	61.7 ±0.2
Gargantua	20.83 ±0.07	38.63 ±0.04	**61.1 ±0.1**
Hunalign	**21.03 ±0.10**	**38.68 ±0.10**	**60.9 ±0.2**
MBA	20.91 ±0.03	**38.85 ±0.14**	61.4 ±0.2

Table 6: SMT results on French–English data.

4. Concluding Remarks

We carried out a brief review of several sentence aligners and evaluated them on the performance of the SMT systems trained on their output, according to automated MT metrics. The coverage of the sentence aligners, as well as the gain achievable by using the best system, depended on the data set. On our Urdu–English data set, this gain was about 1 BLEU point, 0.5 Meteor point and more than 1.5 TER point. On our French–English data set, this gain was about 0.4 BLEU point, 0.3 Meteor point and 0.5 TER point.

Bleualign was the only tool to be ranked first (taking the error range into account) on both tasks and according to the three metrics computed. Gargantua and Hunalign were ranked first according to all metrics on one task. The Gale and Church and MBA tools were ranked first according to one metric on one task.

Acknowledgments

This work has been partially funded by the French Government under the project COSMAT (ANR ANR-09-CORD-004).

5. References

S. Banerjee and A. Lavie. 2005. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. pages 65–72, Ann Arbor, MI, June.

Fabienne Braune and Alexander Fraser. 2010. Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 81–89, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pascale Fung and Percy Cheung. 2004. Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and em. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 57–63, Barcelona, Spain.

William A. Gale and Kenneth W. Church. 1991. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, ACL '91, pages 177–184, Stroudsburg, PA, USA. Association for Computational Linguistics.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Comput. Linguist.*, 19(1):75–102.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Machine Translation Summit X*, pages 79–86.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *AMTA '02: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, pages 135–144, London, UK. Springer-Verlag.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318, Philadelphia, USA, July.

Rico Sennrich and Martin Volk. 2010. MT-based sentence alignment for OCR-generated parallel texts. In *Proceedings of AMTA 2010*, Denver, Colorado.

Rico Sennrich and Martin Volk. 2011. Iterative, MT-based sentence alignment of parallel texts. In *NODALIDA 2011, Nordic Conference of Computational Linguistics*, May.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231, Cambridge, MA, USA.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, and Dan Tufis. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006*, pages 2142–2147.

Dániel Varga, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. 2005. Parallel corpora for medium density languages. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP-2005)*, pages 590–596, Borovets, Bulgaria.

Towards a cross-linguistic VerbNet-style lexicon for Brazilian Portuguese

Carolina Scarton, Sandra Aluísio

Center of Computational Linguistics (NILC), University of São Paulo
Av. Trabalhador São-Carlense, 400. 13560-970 São Carlos/SP, Brazil
carol@icmc.usp.br, sandra@icmc.usp.br

Abstract

This paper presents preliminary results of the Brazilian Portuguese Verbnet (VerbNet.Br). This resource is being built by using other existing Computational Lexical Resources via a semi-automatic method. We identified, automatically, 5688 verbs as candidate members of VerbNet.Br, which are distributed in 257 classes inherited from VerbNet. These preliminary results give us some directions of future work and, since the results were automatically generated, a manual revision of the complete resource is highly desirable.

1. Introduction

The task of building Computational Lexical Resources (CLRs) and making them publicly available is one of the most important tasks of Natural Language Processing (NLP) area. CLRs are used in many other applications in NLP, such as automatic summarization, machine translation and opinion mining. Specially, CLRs that treat the syntactic and semantic behaviour of verbs are very important to the tasks of information retrieval (Croch and King, 2005), semantic parser building (Shi and Mihalcea, 2005), semantic role labeling (Swier and Stevenson, 2004), word sense disambiguation (Girju et al., 2005), and many others. The reason for this is that verbs contain information about sentence roles, such as the argument position, that could be provided by knowing the verb.

The English language has a tradition in building CLRs. The most widely known are WordNet (Fellbaum, 1998), PropBank and its frame files (Palmer et al., 2005), FrameNet (Baker et al., 2005) and VerbNet (Kipper, 2005). All of these resources have information about verbs, but in a different way: WordNet contains deep semantic relations of verbs, such as synonym and hyperonym; PropBank has information about verbs and their arguments with semantic role annotation; FrameNet groups verbs according to the scenario in which these verbs appear; and VerbNet groups verbs according to their syntactic and semantic behaviours. VerbNet-style follows Levin's hypothesis (Levin, 1993), in which verbs that share the same syntactic behaviour also share some semantic components. As an example (from Levin (1993)), let's observe verbs *to spray* and *to load* (sentences 1 and 2).

1. Sharon *sprayed* water on the plants / Sharon *sprayed* the plants with water

2. The farmer *loaded* apples into the cart / The farmer *loaded* the cart with apples

It is possible to see that the verb *to spray* in 1 and *to load* in 2 share the same syntactic behaviour (the objects changed places) and the semantic of these verbs is related to putting and covering something. This alternation of arguments is called diathesis alternation. In this example, it is also possible to see that the semantic of Levin's verb classes is superficial: we can not say that the verb *to spray* is a synonym of the verb *to load*. To fulfill this gap, VerbNet has mappings to WordNet, which has deeper semantic relations.

Brazilian Portuguese language lacks CLRs. There are some initiatives like WordNet.Br (Dias da Silva et al., 2008), that is based on and aligned to WordNet. This resource is the most complete for Brazilian Portuguese language. However, only the verb database is in an advanced stage (it is finished, but without manual validation), currently consisting of 5,860 verbs in 3,713 *synsets*. Other initiatives are PropBank.Br (Duran and Aluisio, 2011), FrameNet.Br (Salomao, 2009) and FrameCorp (Bertoldi and Chishman, 2009). The first one is based on PropBank and the second and third are based on FrameNet.

However, none of these resources tackles the syntactic/semantic interface of the verbs. Therefore, we proposed VerbNet.Br (Scarton, 2011), which is a VerbNet for Brazilian Portuguese language, directly aligned to VerbNet. This is why we started our work from a manual step, which involved manual translation of diathesis alternations of VerbNet from English into Portuguese (see more details in Section 3.1).

Whereas CLRs inspired on WordNet, PropBank and FrameNet have been built by using manual approaches based on corpora, several approaches to build verbnets for other languages employed completely automatic methods, focusing on machine learning. Studies like Joanis and Stevenson (2003), Sun et al. (2008), Sun et al. (2009), Kipper (2005), Merlo and Stevenson (2001) and Sun and Korhonen (2011) for English language, Merlo et al. (2002) for Italian language, Schulte im Walde (2006) for German language, Ferrer (2004) for Spanish language and Sun et al. (2010) for French language focuse on machine learning. Most of these researches used information of frames subcategorization as features for machine learning methods. Subcategorization frames provides information about the syntactic realization of verbs as well as diathesis alternations.

To build VerbNet.Br, we are considering the hypothesis that Levin's verb classes have a cross-linguistic potential - this hypothesis was enunciated by Jackendoff (1990) and verified by Merlo et al. (2002) for Italian, Sun et al. (2010) for French and Kipper (2005) for Portuguese. Using that, we proposed a semi-automatic method to build the VerbNet.Br by using the alignments between WordNet.Br and WordNet and the mappings between VerbNet and WordNet. We also have the hypothesis that this semi-automatic method will present better results (results with more precision) than the completely automatic methods.

In this paper we present the current state of VerbNet.Br project by showing a complete run in the method we have chosen and some preliminary results. In section 2, we present a literature review of CLRs and the relation of these and VerbNet.Br. We also present in this section the relation of VerbNet.Br and some completely automatic methods. In section 3, we present the method to build VerbNet.Br. In section 4, we present preliminary results of VerbNet.Br, using as examples the classes ”Equip-13.4.2”, ”Remove-10.1” and ”Banish-10.2” inherited automatically from VerbNet. Finally, in section 5, we present some conclusions and future work.

2. Literature review

Since VerbNet.Br has been built by using VerbNet, WordNet and WordNet.Br, our literature review is focused on these three resources. Moreover, we also present some completely automatic approaches that are related to our research.

2.1. WordNet

WordNet (Fellbaum, 1998) is the most used CLR. The main semantic relation of this kind of CLR is synonymy - *synsets* are based in this relation. Because of this, Wordnet is composed by four classes: nouns, adjectives, adverbs and verbs (words from different syntactic classes are not synonyms). The verb database contains 11,306 verbs and 13,508 *synsets*.

By using WordNet, wordnets to other languages has been built. MultiWordNet (Bentivogli et al., 2002) and EuroWordNet (Vossen, 2004) are large projects that aim to build wordnets to many other languages such as Italian, Spanish, German, French and Portuguese. WordNet.Br is also based on WordNet.

2.2. WordNet.Br

The Brazilian Portuguese wordnet (called WordNet.Br) (Dias da Silva et al., 2008) is based on WordNet and aligned to it. This CLR is the most complete for Brazilian Portuguese language and has the verb database finished but still under validation. WordNet.Br used the following method:

- A linguist selected a verb in Portuguese;
- Then, he/she searched in a Portuguese-English dictionary for the verb in English that best fitted in the sense in Portuguese;
- After that, he/she searched in WordNet for the *synset* that best fitted in the sense;
- Finally, the linguist decided what kind of relation the *synsets* had. The options were: EQ_SYNONYM (perfect synonym), EQ_NEAR_SYNONYM (imperfect synonym), EQ_HAS_HYPONYM (hyponymy relation) and EQ_HAS_HYPERNYM (hypernymy relation). These relations were defined by Vossen (2004) in the EuroWordNet project.

Figure 1 (from Felippo and Dias da Silva (2007)) shows an example of a *synset* of WordNet aligned to a *synset* of WordNet.Br by using the EQ_SYNONYM alignment.

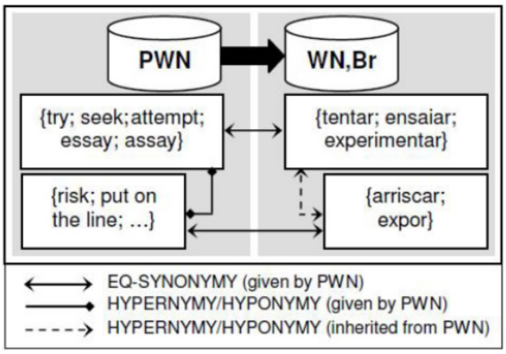

PWN
WN,Br
{try; seek;attempt; essay; assay}
{tentar; ensaiar; experimentar}
{risk; put on the line; ...}
{arriscar; expor}
EQ-SYNONYMY (given by PWN)
HYPERNYMY/HYPONYMY (given by PWN)
HYPERNYMY/HYPONYMY (inherited from PWN)

Figure 1: Example of a *synset* alignment between WordNet and WordNet.Br (Felippo and Dias da Silva, 2007)

As you can see in Figure 1, the other semantic relations, like hypernymy, can be inherited by WordNet.Br from WordNet. This is possible because of the alignment between the *synsets*.

2.3. VerbNet

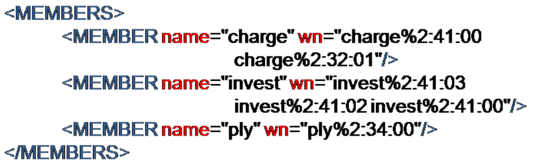
VerbNet (Kipper, 2005) has syntactic and semantic information about English verbs. It is based on Levin's hypothesis of verb classes. This CLR has mappings to PropBank, FrameNet and WordNet.

Verb classes have a group of members, thematic roles, selective restrictions, syntactic frames and semantic predicates. Table 1 shows the structure of ”Equip-13.4.2”, which is a class of VerbNet.

Equip-13.4.2		
Thematic roles and Selectional restrictions: Agent [+animate — +organization], Theme and Recipient [+animate — +organization]		
Members: charge, invest, ply, arm, equip, rearm, redress, regale, reward, saddle, treat, armor, burden, compensate, encumber, overburden, weight		
Frames:		
NP V NP PP	Brown equipped Jones with a camera.	Agent V Recipient with Theme
Semantic Predicates	(1) has_possession(start(E), Agent, Theme); (2) has_possession(end(E), Recipient, Theme); (3) transfer(during(E), Theme); (4) cause(Agent, E)	

Table 1: The structure of ”Equip-13.4.2” class of VerbNet

Each member could be mapped to one or more *synsets* of WordNet, as we can see in Figure 2. The mappings are represented by ”wn” tags.



```
<MEMBERS>
        <MEMBER name="charge" wn="charge%2:41:00
                              charge%2:32:01"/>
        <MEMBER name="invest" wn="invest%2:41:03
                              invest%2:41:02 invest%2:41:00"/>
        <MEMBER name="ply" wn="ply%2:34:00"/>
</MEMBERS>
```

Figure 2: Example of the mappings between VerbNet and WordNet

2.4. Automatic methods

Some studies grouped verbs by using machine learning methods in large corpora. Although the method proposed here is semi-automatic and based on other resources, we also used some techniques of these studies and we intend to compare the results of our method with the results of a machine learning method.

For the English language, studies of Joanis and Stevenson (2003), Merlo and Stevenson (2001), Kipper (2005), Sun et al. (2008) and Sun et al. (2009) presented methods to group verbs automatically. Especially, Kipper (2005) made experiments with machine learning to improve the VerbNet. Sun et al. (2008), Sun et al. (2009) and Joanis and Stevenson (2003) considered the Levin's taxonomy to put verbs into classes.

For other languages, we can cite Sun et al. (2010) (French), Ferrer (2004) (Spanish), Merlo et al. (2002) (Italian) and Schulte im Walde (2006) (German). Specifically, Sun et al. (2010) used a gold standard to compare with the machine learning results. The building of this gold standard was quite similar to our method to build VerbNet.Br. Besides that, Sun et al. (2010), Merlo et al. (2002) and Schulte im Walde (2006) also considered the Levin's taxonomy.

Most of these researches used subcategorization frames as features for machine learning. In our approach, we use subcategorization frames too, but in a different way (see Section 3). However, we also intend to evaluate the results of our semi-automatic method, comparing them with the results of a completely automatic method that will use machine learning with subcategorization frames as features.

3. Building VerbNet.Br

Although Scarton (2011) reported the method developed to build the VerbNet.Br, such paper is available only in Portuguese and, for this reason, we decided to quickly describe it here. The proposed method is composed by four stages (Sections 3.1, 3.2, 3.3 and 3.4, respectively, present the four stages). We based our experiments on version 3.0 of VerbNet and we only considered the classes defined by Levin (1993) without the subclasses and extensions proposed by Kipper (2005).

3.1. Stage 1: Manual translation of diathesis alternations of VerbNet from English into Portuguese

The Stage 1 (under development) is the direct translation of diathesis alternations from English into Portuguese, manually. For example, Table 1 presents only one diathesis alternation for the class "Equip-13.4.2": "NP V NP with NP", that means, a noun phrase followed by a verb, followed by a noun phrase, followed by the preposition "with", followed by a noun phrase. This alternation can be directly translated into Portuguese:"NP V NP *com* NP". To do that, we just replaced the preposition "with" in English for the preposition *com* in Portuguese. In this step, we only consider the alternations that can be directly translated. If an alternation doesn't occur in Portuguese or if it occurs in a different way, it is not translated.

We decided to translate only the alternations that fits perfectly into Portuguese because of two reasons. The first one is that we did not have specialized people to do this task. The task is being developed by a native speaker of Portuguese, who does not have linguistic expertise. The second one is that we intend to identify the similarity between English and Portuguese diathesis alternations and find out how many diathesis alternations are shared by both languages. Besides that, we intend firstly to establish the perfect alignments and, after, deal with the other cases. As future work, we intend to extend VerbNet.Br with alternations that were not directly translated and with alternations that appear in Portuguese, but not in English, such as phrases without subject.

3.2. Stage 2: Automatic search of diathesis alternations of Brazilian Portuguese verbs in corpus

The Stage 2 (finished) is the search for diathesis alternations of verbs in Portuguese in corpus. This step was carried out by using the subcategorization frames extractor tool developed by Zanette (2010). This tool, based on Messiant (2008) developed for the French language, uses a corpus, tagged by PALAVRAS parser (Bick, 2005), to identify the syntactic behaviour of verbs. In other words, the search was for patterns like "NP V NP", "NP V com NP", etc (Zanette et al., 2012).

The Lácio-ref (Aluísio et al., 2004), a Brazilian Portuguese corpus from Lácio-Web project, was used in this stage. This corpus has, approximately, 9 million words and it is divided into five genres: scientific, informative, law, literary, and instructional. We identified 8,926 verbs and 196,130 frames. However, these numbers also contain incorrect verbs and incorrect frames that will be discarded by using a threshold frequency.

For example, the verbs of class "Equip-13.4.2" should present in the corpus the pattern "NP V NP *com* NP" as defined in the Stage 1.

3.3. Stage 3: Automatic generation of candidate members of VerbNet.Br by using other CLRs

The Stage 3 (finished) was the generation of candidate members for classes of VerbNet.Br, by using the mappings between VerbNet and WordNet and the alignments between WordNet and WordNet.Br. Figure 3 shows how this stage was developed: for each class in VerbNet, we searched firstly the *synsets* of WordNet mapped to each verb member, then we searched for the *synsets* of WordNet.Br and thus the members of these Portuguese *synsets* were defined as the candidate members. We defined 4,063 verbs as candidate members in 207 classes.

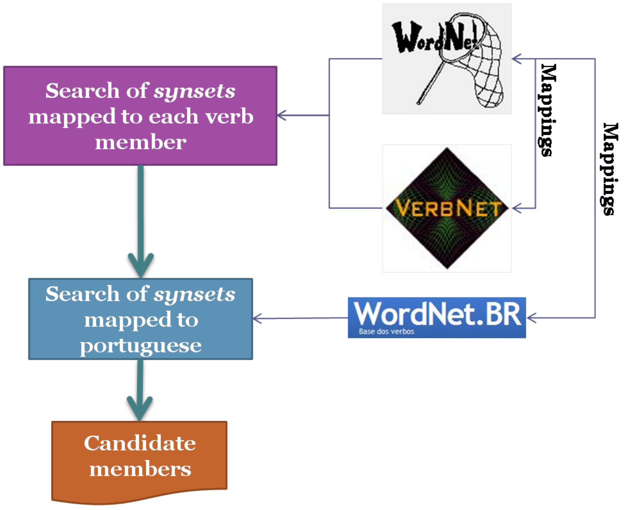

Search of synsets mapped to each verb member
WordNet
VERBNET
Mappings
Mappings
Search of synsets mapped to portuguese
WordNet.BR
Base dos verbos
Candidate members

Figure 3: Candidate Members definition

For the class ”Equip-13.4.2” we identified 38 candidate members, such as *dotar* (to gift) and *armar* (to arm).

3.4. Stage 4: Selection of members of VerbNet.Br CLRs

Finally, the Stage 4 (future work) will use all the others together. Figure 4 shows an illustration of how this stage will work.

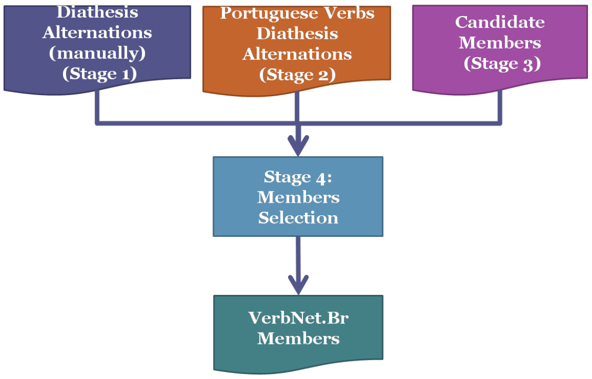

Diathesis Alternations (manually) (Stage 1)
Portuguese Verbs Diathesis Alternations (Stage 2)
Candidate Members (Stage 3)
Stage 4: Members Selection
VerbNet.Br Members

Figure 4: Stage 4: putting all stages together

As may be seen in Figure 4, this stage will use all the other stages to select the members of VerbNet.Br. For each candidate member, we will compare the diathesis alternations identified in the Stage 2 with the diathesis alternations translated in the Stage 1. If the candidate member presents in the corpus (Stage 2) a certain frequency of the diathesis alternations defined in Stage 1, it will be selected, if not, it will be discarded. Some results of this stage, from a pilot test, will be presented in the next section.

4. Experiments

This section contains the preliminary results of VerbNet.Br. Since the Stages 2 and 3 are already done, we carried out an experiment with three classes taken from the Stage 1. The classes selected were ”Equip-13.4.2”, which is shown in Table 1, ”Remove-10.1”, shown in Table 2, and ”Banish-10.2”, shown in Table 3.

Remove-10.1		
Thematic roles and Selectional restrictions: Agent [+int_control — +organization], Theme and Source [+location]		
Members: abolish, abstract, cull, deduct, delete, depose, disgorge, dislodge, disengage, draw, eject, eliminate, eradicate, excise, excommunicate, expel, extirpate, extract, extrude, lop, omit, ostracize, partition, pry, reap, retract, roust, separate, shoo, subtract, uproot, winkle, wrench, withdraw, oust, discharge, dismiss, evict, remove, sever		
Frames:		
NP V NP	Doug removed the smudges.	Agent V Theme
Semantic Predicates	(1) cause(Agent, E) (2) location(start(E), Theme, ?Source) (3) not(location(end(E), Theme, ?Source))	
NP V NP PP.source	Doug removed the smudges from the tabletop.	Agent V Theme +src Source
Semantic Predicates	(1) cause(Agent, E); (2) location(start(E), Theme, Source); (3) not(location(end(E), Theme, Source))	

Table 2: The structure of ”Remove-10.1” class of VerbNet

Section 4.1 presents materials and methods. Section 4.2 contains the preliminary results for the three classes cited above.

4.1. Materials and methods

Since the Stages 2 and 3 are stored in a MySQL database, it was easy to recover the data and to compare it. The Stage 1 is being developed in XML files, making automatic information recovery easy too.

The subcategorization frames identified in Stage 2 needed to be filtered out mainly because of some parsing errors like adjuncts tagged as arguments. Therefore, the Maximum Likelihood Estimate (MLE), used in previous work (Ferrer, 2004), was applied in this phase. The MLE is the ratio of the frequency of a verb frame to the whole frequency of the verb. We considered a threshold of 0,05 (the same adopted by Ferrer (2004)).

We also needed to decide how many diathesis alternations we would consider to select a candidate member. For these preliminary experiments, the rate of 60% was our choice, although we will also test other values. This was important because some diathesis alternations defined in the Stage 1 did not occur in the corpus (the alternation could be easily and correctly generated, but they were never used by native speakers). The rate of 60% was chosen empirically. As future work, we intend to vary this rate (50%, 70%, etc) and to evaluate the impact of this rate in the overall precision and recall.

Banish-10.2		
Thematic roles and Selectional restrictions: Agent [+animate — +organization], Theme[+animate], Source [+location] and Destination [+location — -region]		
Members: banish, deport, evacuate, expel, extradite, recall, remove, shanghai		
Frames:		
NP V NP	The king banished the general.	Agent V Theme
Semantic Predicates	(1) cause(Agent, E); (2) location(start(E), Theme, ?Source); (3) location(end(E), Theme, ?Destination)	
NP V NP PP.source	The king banished the general from the army.	Agent V Theme +src Source
Semantic Predicates	(1) cause(Agent, E); (2) location(start(E), Theme, Source); (3) not(location(end(E), Theme, Source))	
NP V NP PP.destination	The king deported the general to the isle.	Agent V Theme to Destination
Semantic Predicates	(1) cause(Agent, E); (2) location(start(E), Theme, ?Source); (3) location(end(E), Theme, Destination)	

Table 3: The structure of ”Banish-10.2” class of VerbNet

4.1.1. Preliminary Results

In this section we present some preliminary results of VerbNet.Br, by using the classes ”Equip-13.4.2”, ”Banish-10.2” and ”Remove-10.1”.

Equip-13.4.2

The class ”Equip-13.4.2” has only one syntactic frame: ”NP V NP with NP” (as shown in Table 1). In the Stage 1, this frame was directly translated into Portuguese: ”NP V NP *com* NP”. Since we have only one syntactic frame, we selected it to be the parameter to discard or to select a candidate member.

In the Stage 3, 38 candidate members were defined for the class ”Equip-13.4.2”. Searching in the results of Stage 2, only 12 verbs presented the syntactic frame defined in the Stage 1. However, only the verb *dotar* (to gift) presented a threshold higher than 0,05. Therefore, the Portuguese version of the class ”Equip-13.4.2” has one syntactic frame (as defined above) and only one member: the verb *dotar* (to gift). In order to verify if the verb *dotar* (to gift) was correctly selected, we evaluated the sentences in the corpus from which the syntactic frame was derived. Two sentences were found:

1. *A natureza dotara Aurélia com a inteligência viva e brilhante[...]* (Nature gifted Aurélia with a bright, vibrant intelligence.)
2. *Era tão universal e inventivo, que dotou a poesia malaia com um novo metro[...]* (He was so universal and creative that he has gifted malayan poetry with a new meter.).

The two sentences present the semantic of the class: X gives something to Y that Y needs. However, if we go back to the Table 1, some of the requirements are missed. For example, the first argument needs to be an animate Agent or an organization and in the first sentence the first argument (*A natureza* - Nature) is not animate neither an organization. This may be explained because Nature was used in a figurative way and took the place of an animate entity. This class is shown in Table 4.

Equip-13.4.2 - BR		
Thematic roles and Selectional restrictions: Agent [+animate — +organization], Theme and Recipient [+animate — +organization]		
Members: *dotar* (to gift)		
Frames:		
NP V NP PP	*Brown dotou Jones com uma câmera.*	Agent V Recipient com Theme
Semantic Predicates	(1) has_possession(start(E), Agent, Theme); (2) has_possession(end(E), Recipient, Theme); (3) transfer(during(E), Theme); (4) cause(Agent, E)	

Table 4: The structure of ”Equip-13.4.2” class of VerbNet.Br

Remove-10.1

Finally, for the class Remove-10.1, the two diathesis alternations (shown in Table 2) were translated from English into Portuguese: ”NP V NP” and ”NP V NP *de* NP”. To be a member, a verb needed to present two of these syntactic frames (the roof of 1.2 (0.6*2)), respecting the MLE measure.

In Stage 3, 151 verbs were identified. Looking at the results from Stage 2 , we found 85 verbs that present at least one of the syntactic frames. Selecting only verbs that present the two diathesis alternations defined for this class by using the threshold of 0.05, we found the verbs *arredar* (to move away), *destituir* (to oust), *diminuir* (to decrease), *dispensar* (to dismiss), *excluir* (to exclude), *isolar* (to isolate), *separar* (to separate) and *tirar* (to remove). Searching for sentences of verb *separar* (to separate) we found two examples:

1. *O vaqueiro separa escrupulosamente a grande maioria de novas cabeas pertencentes ao patrão[...]* (The cowboy carefully picks out most of the new cattle belonging to his master.)
2. *Cetonas em estado de triplete podem separar hidrogênios de grupos benzilas[...]* (Ketones in triplet states can separate hydrogen from benzyl groups.)

The semantic of this class is ”the removal of an entity from a location” (Levin, 1993). The sentences presented before follow this semantic and respect the restrictions defined for the thematic roles (shown in Table 2). This class is presented in Table 5.

Remove-10.1 - BR		
Thematic roles and Selectional restrictions: Agent [+int_control — +organization], Theme and Source [+location]		
Members: *arredar* (to move away), *destituir* (to oust), *diminuir* (to decrease), *dispensar* (to dismiss), *excluir* (to exclude), *isolar* (to isolate), *separar* (to separate) and *tirar* (to remove)		
Frames:		
NP V NP	Doug removeu as manchas.	Agent V Theme
Semantic Predicates	(1) cause(Agent, E) (2) location(start(E), Theme, ?Source) (3) not(location(end(E), Theme, ?Source))	
NP V NP PP.source	Doug removeu as manchas da toalha.	Agent V Theme +src Source
Semantic Predicates	(1) cause(Agent, E); (2) location(start(E), Theme, Source); (3) not(location(end(E), Theme, Source))	

Table 5: The structure of ”Remove-10.1” class of VerbNet.Br

Banish-10.2

The class ”Banish-10.2” has three syntactic frames (as shown in Table 3). In the Stage 1, we translated directly all of these: ”NP V NP”, ”NP V NP *de* NP” and ”NP V NP *para* NP”. To be a member, a verb needed to present two of these syntactic frames (the roof of 1.8 (0.6*3)), respecting the MLE measure.

In the Stage 3, 35 verbs were defined for this class. Searching in the results of the Stage 2, we found 18 verbs that present at least one of the syntactic frames. However only the verbs *excluir* (to exclude) and *tirar* (to remove) present at least 2 syntactic frames that have a threshold higher than 0.05. Both presented the same syntactic frames: NP V NP and NP V NP *de* NP.

Therefore, the Portuguese version of the class ”Banish-10.2” has two verbs, *excluir* (to exclude) and *tirar* (to remove), and presents two syntactic frames: NP V NP and NP V NP *de* NP. Searching for sentences of the verb *excluir* (to exclude), we found two examples:

1. *[...] outras espécies [...] excluem as espécies responsáveis pela mudança.* (Other species exclude the species responsible for the change.)

The semantic of this class is ”removal of an entity, typically a person, from a location” (Levin, 1993). The sentence presented fits in this semantics, but we could not find an example of the alternation ”NP V NP *de* NP” with the second NP (Theme) being animate. We only find sentences that fit in the semantic of Remove-10.1 class. This class is shown in Table 6 (the ? means that the sentence seems to be incorrect, according to the corpus we have used).

Banish-10.2 - BR		
Thematic roles and Selectional restrictions: Agent [+animate — +organization], Theme, Source [+location] and Destination [+location — -region]		
Members: *excluir* (to exclude) and *tirar* (to remove)		
Frames:		
NP V NP	*O rei excluiu o general.*	Agent V Theme
Semantic Predicates	(1) cause(Agent, E); (2) location(start(E), Theme, ?Source); (3) location(end(E), Theme, ?Destination)	
NP V NP PP.source	*O rei excluiu o general do exército.*	Agent V Theme +src Source
Semantic Predicates	(1) cause(Agent, E); (2) location(start(E), Theme, Source); (3) not(location(end(E), Theme, Source))	
NP V NP PP.destination	?*O rei excluiu o general para a ilha.*	Agent V Theme para Destination
Semantic Predicates	(1) cause(Agent, E); (2) location(start(E), Theme, ?Source); (3) location(end(E), Theme, Destination)	

Table 6: The structure of ”Banish-10.2” class of VerbNet.Br

5. Conclusions and Future Work

We have presented a semi-automatic method for building the VerbNet.Br and some preliminary results with three classes. The classes presented were ”Equip-13.4.2”, ”Remove-10.1” and ”Banish-10.2”. The second and the third ones are related, since they have almost the same meaning and differ only in some diathesis alternations. The thematic roles, selectional restrictions and semantic predicates will be directly inherited from English. As the proposed method uses existing resources in one language to build a new resource in another language, it is cross-linguistic, that is, the method explores the compatibilities between English and Portuguese languages. However, we can observe that a linguistic revision of the results of this semi-automatic method is highly desirable. Therefore, we are looking for collaborators interested in validating this resource.

As future work, we intend to finish stages one and four and apply the method for all the remaining classes. We will also change the thresholds used to evaluate the precision and recall. Besides that, we will evaluate how many verbs are defined as candidate members (result of Stage 3) and how many verbs are selected (result of Stage 4). This will be achieved by calculating the ratio of selected verbs to candidate verbs.

We will also use a completely automatic method to group verbs, by using machine learning. This method will use clustering to group verbs according to subcatecategorization frames. We intend to compare the resulting classes of this automatic method with classes of our semi-automatic method proposed. We have the hypothesis that the semi-automatic method will present classes with more precision. However, the automatic method is expected to have a best recall.

Since we expect that the automatic method will present more verbs, we will try to include these verbs in VerbNet.Br classes and improve the resource, similarly to the task carried out by Kipper (2005).

6. Acknowledgements

We thank the Brazilian Science Foundation FAPESP for financial support.

7. References

S. M. Aluísio, G. M. Pinheiro, A. M. P. Manfrim, L. H. M. Genovês Jr., and S. E. O. Tagnin. 2004. The Lácio-web: Corpora and Tools to Advance Brazilian Portuguese Language Investigations and Computational Linguistic Tools. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1779–1782, Lisbon, Portugal.

C. F. Baker, C. J. Fillmore, and J. F. Lowe. 2005. The Berkeley Framenet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 86–90, University of Montréal, Canadá.

L. Bentivogli, E. Pianta, and C. Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet Conference*, pages 293–302, Mysore, India.

A. Bertoldi and R. L. O. Chishman. 2009. Desafios para a Criação de um Léxico baseado em Frames para o Português: um estudo dos frames Judgment e Assessing. In *Proceedings of the The 7th Brazilian Symposium in Information and Human Language Technology*, São Carlos, SP, Brazil.

E. Bick. 2005. *The Parsing System*. Ph.D. thesis.

D. Croch and T. H. King. 2005. Unifying Lexical Resources. In *Proceedings of Interdisciplinary Workshop on the Identication and Representation of Verb Features and Verb*, pages 32–37, Saarbruecken, Germany.

B. C. Dias da Silva, A. Di Felippo, and M. G. V. Nunes. 2008. The Automatic Mapping of Princeton Wordnet lexical-conceptual relations onto the Brazilian Portuguese Wordnet database. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 1535–1541, Marrakech, Morocco.

M. S. Duran and S. M. Aluisio. 2011. Propbank-br: a Brazilian Portuguese corpus annotated with semantic role labels. In *Proceedings of The 8th Brazilian Symposium in Information and Human Language Technology*, Cuiabá, MT, Brazil.

A. Di Felippo and B. C. Dias da Silva. 2007. Towards na automatic strategy for acquiring the Wordnet.br hierarchical relations. In *Proceedings of the 5th Workshop in Information and Human Language Technology, in conjunction with XXVII Congresso da Sociedade Brasileira de Computao*, pages 1717–1720, Rio de Janeiro, RJ, Brazil.

C. Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT Press, Cambridge, MA.

E. E. Ferrer. 2004. Towards a semantic classification of spanish verbs based on subcategorisation information. In *Proceedings of the Workshop on Student research, in conjunction with ACL 2004*, pages 163–170, Barcelona, Spain.

R. Girju, D. Roth, and M. Sammons. 2005. Token-level Disambiguation of Verbnet Classes. In *Proceedings of Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, Saarbruecken, Germany.

R. Jackendoff. 1990. *Semantic Structures*. MIT Press, Cambridge, MA.

E. Joanis and S. Stevenson. 2003. A general feature space for automatic verb classification. In *Proceedings of the 10th conference on European chapter of the Association for Computational Linguistics*, pages 163–170, Budapest, Hungria.

K. Kipper. 2005. *Verbnet: A broad coverage, comprehensive verb lexicon.* Doctor of philosophy, University of Pennsylvania.

B. Levin. 1993. *English Verb Classes and Alternation, A Preliminary Investigation*. The University of Chicago Press, Chicago, IL.

P. Merlo and S. Stevenson. 2001. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27(3):373–408.

P. Merlo, S. Stevenson, V. Tsang, and G. Allaria. 2002. A multilingual paradigm for automatic verb classification. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 207–214, Philadelphia, PA.

C. Messiant. 2008. A subcategorization acquisition system for French verbs. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Techonologies*, pages 55–60, Columbus,OH.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The Proposition Bank: A Corpus Annotated with Semantic Roles. *Computational Linguistics*, 31(1):71–106.

M. M. Salomao. 2009. Framenet Brasil: Um trabalho em progresso. *Revista Calidoscópio*, 7(3):171–182.

C. Scarton. 2011. Verbnet.br: construção semiautomática de um léxico computacional de verbos para o Português do Brasil. In *Proceedings of the The 8th Brazilian Sym-*

posium in Information and Human Language Technology, Cuiabá, MT, Brazil.

S. Schulte im Walde. 2006. Experiments on the Automatic Induction of German Semantic Verb Classes. *Computational Linguistics*, 32(2):159–194.

L. Shi and R. Mihalcea. 2005. Putting Pieces Together: Combining Framenet, Verbnet and Wordnet for Robust Semantic Parsing. In *Proceedings of 6th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 99–110, Mexico City, Mexico.

L. Sun and A. Korhonen. 2011. Hierarchical Verb Clustering Using Graph Factorization. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1023–1033, Edinburgh, UK.

L. Sun, A. Korhonen, and Y. Krymolowski. 2008. Verb class discovery from rich syntactic data. In *Proceedings of the 9th international conference on Computational linguistics and intelligent text processing*, pages 16–27, Haifa, Israel.

L. Sun, A. Korhonen, and Y. Krymolowski. 2009. Improving verb clustering with automatically acquired selectional preferences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 638–647, Singapore.

L. Sun, A. Korhonen, T. Poibeau, and C. Messiant. 2010. Investigating the cross-linguistic potential of Verbnet-style classification. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1056–1064, Beijing, China.

R. Swier and S. Stevenson. 2004. Unsupervised Semantic Role Labelling. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 95–102, Barcelona, Spain.

P. Vossen. 2004. Eurowordnet: a multilingual database of autonomous and language specific wordnets connected via an interlingual-index. *International Journal of Linguistics*, 17.

A. Zanette, C. Scarton, and L. Zilio. 2012. Automatic extraction of subcategorization frames from corpora: an approach to Portuguese. In *Proceedings of the 2012 International Conference on Computational Processing of Portuguese - Demo Session*, Coimbra, Portugal.

A. Zanette. 2010. *Aquisiçao de Subcategorization Frames para Verbos da Língua Portuguesa.* Projeto de diplomação, Federal University of Rio Grande do Sul.

Multiword Named Entities Extraction from Cross-Language Text Re-use

Parth Gupta[1], **Khushboo Singhal**[2], **Paolo Rosso**[1]

[1] Natural Language Engineering Lab - ELiRF
Department of Information Systems and Computation
Universidad Politécnica de Valencia, Spain
http://users.dsic.upv.es/grupos/nle
{pgupta,prosso}@dsic.upv.es

[2] IR-Lab, DA-IICT, India.
http://irlab.daiict.ac.in
khushboo_singhal@daiict.ac.in

Abstract

In practice, many named entities (NEs) are multiword. Most of the research, done on mining the NEs from the comparable corpora, is focused on the single word transliterated NEs. This work presents an approach to mine Multiword Named Entities (MWNEs) from the text re-use document pairs. Text re-use, at document level, can be seen as noisy parallel or comparable text based on the level of obfuscation. Results, reported for Hindi-English language pair, are very encouraging. The approach can easily be extended to any language pair.

1. Introduction

Text re-use refers to using the text again from its original source. There are different situations which fall under the category of text re-use e.g. paraphrasing, quotation and copying (plagiarism). Moreover, text re-use is not limited to a single language, it can be cross-lingual in case of translated documents and cross-language plagiarism. Detection of such text re-use helps in various applications, e.g. checking the authenticity of the text, identifying near duplicates. Moreover, the identified document pairs can also be exploited for mining natural language resources. The difficulty of detection of re-use even increases when the source and target texts are in different languages which is called cross-language text re-use. There are two levels of text re-use:

1. Document level: The entire text of the document is re-used from some source, and
2. Fragment level: One or some of the sections of the document are containing re-used text.

Irrespective of the types and levels of the re-use, both, the source and target texts talk about the same concept with a high overlap in semantics and paraphrasing compared to an independent original work on the same topic. From now onward, we would talk in context of cross-language text re-use which can also be seen as noisy parallel or comparable text based on the level of obfuscation. This makes it more exploitable for mining the various cross-language resources like named entities, multiwords expression units, translation and transliteration probabilities.

Multiword units are very useful in many natural language processing (NLP) applications like multiword expressions for phrase based statistical machine translation (SMT) (Lambert and Banchs, 2005), MWNEs for cross-language news aggregation, finding NE equivalents in multilingual environment and measuring cross-language similarity for finding potential near-translation of the document from the multilingual corpora (Steinberger et al., 2002).

Named entities are very efficient elements in the cross language information retrieval (CLIR) and NLP applications like machine translation, machine transliteration, mention detection (Zitouni and Florian, 2008), news aggregation (Liu and Birnbaum, 2008) and plagiarism detection (Gupta et al., 2010). There have been many approaches for machine transliteration in order to find and use NEs in respective applications (Karimi et al., 2011). As suggested by Oh et al. (2006), the transliterations generated by the statistical methods are not often accurate, moreover, there can be more than one transliterations possible for a particular term. Therefore, it makes more sense to *mine* the NEs from the readily available multilingual resources like parallel and comparable text. On a similar note, Udupa et al. (2008) and Klementiev and Roth (2006), both, attempt to mine NEs from a comparable multilingual corpora. A considerable amount of research has been done on the extraction of NEs from the comparable corpora, but most of the methods at the core are meant for the single word NEs and more specifically transliterated single word NEs. Bhole et al. (2011) suggested an approach to mine MWNEs from a comparable corpus, which can be seen as very close to the approach we propose in this paper. The key difference lies in the prior knowledge and the problem formulation, the former tries to formulate the problem as a conditional probability of target language MWNE alignment for the given source language MWNE, while we do not assume any prior knowledge of source language MWNE and pose the problem as joint probability estimation.

Though *enough* mono-lingual, and to some extent cross-lingual, resources are available for Hindi-English, they are still not abundant to solve the general problems of NLP with high accuracy, compared to that achieved for some peer English language pairs e.g. English-Spanish. This lag is due to the absence of sufficient parallel data and, to an extent, technological and cultural inadequacy for Hindi resource creation environment e.g. (less Hindi speakers prefer to use computers in Hindi and even less people use a Hindi keyboard). This makes it more important to exploit the present *poor* resources to the fullest.

The rest of the paper is structured as follows. Section 2 talks about the importance and challenges involved with MWNE. Section 3 presents the proposed approach in detail. In Section 4 we describe the corpus and report results with analysis. Finally in Section 5, we conclude the work with some future directions.

2. Motivation

A general observation gives an insight to the nature of NEs that many NEs are multiword e.g. name of a person with the surname (e.g. Barrak Obama), full name of an organization (e.g. Technical University of Valencia), city name with the country name (e.g. Valencia, Spain) and so on.

In order to understand the distribution and amount of the MWNEs, we tagged 2275 English news articles[1] using an English NE recogniser[2] (NER). Out of total 21,208 unique NEs: 9,079 (43%) were single word and 12,129 (57%) were multi-word NEs. This demonstrates the importance for explicit handling of the MWNEs.

Bhole et al. (2011) report the issues involved in finding the MWNEs and the nature of MWNEs. MWNEs are not merely a transliteration of terms, rather they may include translation, sequential shuffling, acronyms, one-to-many and many-to-one correspondence among the terms and so on.

The limitation of the conditional probability estimation based method is that the performance is dependent on the accuracy and efficiency of the source language NE recognition. To understand this phenomenon, we carried an experiment where we tagged the English documents using an English NER. We noticed that the NER identifies some of the NEs partly, for example "Bayes" instead of "Bayes Theorem", "Sundereshwara Temple" instead of "Meenakshi Sundereshwara Temple". In addition, there were many false positives. Finding the MWNEs in target language based on these source MWNEs will lead to a very noisy identification which needs to be handled by pruning. Therefore, we pose the problem of MWNE identification as the estimation of a joint probability for two string sequences being an MWNE pair.

3. Algorithm

First, the text re-use document pairs from the non-comparable source collection are found based on the standard CLIR methods of query translation. We consider the Hindi document as the query and retrieve the most similar English document from the indexed source collection. After fetching such pairs, we mine them to extract the MWNEs. For finding re-used document pairs, we use the system reported in (Gupta and Singhal, 2011).

3.1. Multiword Named Entity Extraction

First of all we find the transliteration match between the source and the target document. Suppose the terms s_{match} and t_{match} represent the corresponding terms of the transliteration match in the source and target documents respectively. Let $S = \{s_1, \cdots, s_N\}$ be a multiword unit including and around s_{match} of the source language (English) of length N and similarly, let $T = \{t_1, \cdots, t_M\}$ be the target language (Hindi) multiword unit including and around t_{match} of length M. The multiword pair <S,T> which maximises the Eq. 2 as shown below, is considered as MWNE.

$$\begin{aligned} &\max \zeta(S,T) \text{ subject to} \qquad (1)\\ &\qquad \phi(S,T) = \min(N,M)\ ,\\ &\qquad |N-M| \leq 1 \text{ and}\\ &\qquad \sum_{i=1}^{N}\sum_{j=1}^{M} T(s_i,t_j) \geq \theta \end{aligned}$$

where,

$$\zeta(S,T) = \ell_s(S) * \ell_t(T) * \sum_{i=1}^{N}\sum_{j=1}^{M} \delta(s_i,t_j) \qquad (2)$$

$$\phi(S,T) = \sum_{i=1}^{N}\sum_{j=1}^{M} \psi(s_i,t_j) \qquad (3)$$

$$\delta(s_i,t_j) = \begin{cases} D(s_i,t_j) & \text{if translation}\\ T(s_i,t_j) & \text{if transliteration}\\ 0 & \text{otherwise} \end{cases} \qquad (4)$$

$$\psi(s_i,t_j) = 1 \qquad \text{if } \delta(s_i,t_j) \neq 0 \qquad (5)$$

ℓ_s and ℓ_t define source and target language model respectively. $D(s_i,t_i) \neq 0$ when dictionary translation for term s_i is t_j, and accordingly $T(s_i,t_j)$ signifies the same for transliteration engine. We consider it an exception and ignore the term s_i when $D(s_i,t_j) = T(s_i,t_j)$, i.e. the translation and transliteration of term s_i is the term t_j. The values taken by $\delta(s_i,t_j)$ are normalised values. For a multiword unit to be a NE, at least one of its terms has to be a transliteration which is maintained by assigning the third condition in Eq. (1) where θ can be set accordingly.

Basically s_{match} and t_{match} help to locate the area of the document pair where the chances of finding an NE is very high. Then after, the approach selects the longest substring pair around s_{match} and t_{match} to be an MWNE pair using the above formulation.

4. Results and Analysis

We report the results of our proposed algorithm on the recently developed corpus called CL!TR-2011 which contains the cross-language text re-use documents of Hindi and English.

[1] Articles are taken from the year 2007 crawl of "The Telegraph" of section "frontpage", which can also be accessed through English Corpus of Forum for Information Retrieval Evaluation (FIRE) 2010.

[2] Alias-i. 2008. LingPipe 4.1.0. We use `cmd_ne_en_news_muc6.sh` script. `http://alias-i.com/lingpipe` (accessed June 25, 2010)

4.1. Corpus

The CL!TR-2011-Dataset[3] contains 190 Hindi documents and 5032 English documents. The documents in the corpus are generated from Wikipedia[4] and are related to the "tourism" and "computer science" domains. Table 1 contains the basic information about the corpus in terms of the size. More details about the corpus can be found in (Barrón-Cedeño et al., 2011).

Partition	$\lvert D\rvert$	$\lvert D_{tokens}\rvert$	$\lvert D_{types}\rvert$
D_{hi}	388	216 k	5 k
D_{en}	5032	9.3 M	644 k

Table 1: Statistics of the CL!TR-2011-Dataset. D_{hi} represents the Hindi document set and D_{en} represents the English document set. $|.|$ is the $size - of$ function.

The corpus contains four types of Hindi documents, mainly categorized by the amount of obfuscation of re-use, namely "Exact", "Heavy", "Light" and "None". The text re-use is through the machine translation with manual corrections. The "Exact" refers to the exact re-use of the text without any modifications, "Heavy" refers to the re-use with very less modifications, "Light" refers to the re-use with high modifications and "None" refers to no re-use.

4.2. Evaluation

Tables 2 and 3 summarize the performance of the text re-use document pair finding module. The configuration (morphological analyser (M) + bilingual dictionary (D) + transliteration (T)), which produced the best results, is used to retrieve the text re-use pair.

Method	**Precision**	**Recall**	**F-Measure**
M+D+T	0.695	0.904	0.786

Table 2: Performance of finding text re-use document pairs on test data. M+D+T signifies the combination of morphological analyser, bilingual dictionary and transliteration for query translation.

Type	Exact	Heavy	Light
Recall	1.0000	0.9070	0.8551

Table 3: Performance evaluation based on different levels of re-use.

In order to extract the MWNEs from the identified document pairs, we give the re-use document pairs of type "Exact", which are 34 in total, to the MWNE extraction module. For the evaluation of MWNE extraction module, we manually identify MWNE pairs in these 34 re-used document pairs, which serves as the gold standard. We limit our evaluation to only type "Exact" because in this preliminary study we wish to investigate the behaviour of the algorithm in a smaller controlled environment.

4.3. Analysis

We use the Universal Word (UW)[5] Version 3.1 Hindi-English bilingual dictionary to represent the $D(s_i, t_j)$ and use Google Transliterate API[6] to represent the $T(s_i, t_j)$ in Eq. (4). The language model for the source and the target languages are computed on the respective language subsets of the CL!TR-2011-Dataset. Results obtained for the MWNE extraction algorithm, are reported in Table 4. We consider two types of results, full match (FM): where a complete MWNE is identified and, partial match (PM): where a part of the MWNE is identified.

Type	**Precision**	**Recall**	**F-Measure**
FM	0.57	0.38	0.49
FM+PM	0.86	0.57	0.69

Table 4: Performance evaluation of MWNE extraction algorithm on Hindi-English language pair. FM is full match and FM+PM is full and partial match.

The corpus contains some of the documents related to "Computer Science" domain, which have some small scientific notations in the text, such as, $P(b|a)$. These notations are identified by the algorithm as an MWNE, and hence hurt the precision. Some examples of this phenomenon along with other false positives are depicted in 7. We take the transliteration engine as a binary model i.e the exact transliteration is considered, though the algorithm is capable to handle continuous values. Hence, near transliterations are missed, which in turn, hurts the recall. The reported results are for multiword NEs and we do not consider single-word NEs for the experimentation. $\phi(S, T)$ in Eq. (1) keeps an account of the number of term pairs contributing to the final score, applying this as a condition in Eq. (1) helps to determine the boundary of the MWNE. The language model helps in voting out the false positives in terms of unnecessary translation match, which is not a part of the MWNE, for example, in case of (English: "Indo Aryan and", Hindi: "ईन्डो आर्यन और") the trailing "the" is removed providing the tighter boundaries.

English	**Hindi**
Sawai Madhopur	स्वाई माधोपुर्
DSIR model	डीएसआईआर मॉडल
Government of Madras	मद्रास सरकार
Kashgar Ladakh	कशगर लद्दाख
Medical Board	मेडिकल बोर्ड

Table 5: Examples of correctly identified full MWNEs.

English	**Hindi**
Ranthambore National Park	रणथंभोर राष्ट्रीय
Meenakshi Amman	मीनाक्शी अम्मान
India Company	इन्डिया कंपनी
administered Gilgit	प्रशासित गिलगित

Table 6: Examples of partially identified MWNEs.

[3] `http://users.dsic.upv.es/grupos/nle/fire-workshop-clitr.html`

[4] `http://www.wikipedia.org/`

[5] `http://www.cfilt.iitb.ac.in/~hdict/webinterface_user/index.php`

[6] `http://code.google.com/apis/language/`

English	Hindi
computing these values	मूल्यों कंप्यूटिंग
year the Sikhs	वर्ष सिखों
probability of b	प्रयिकता बी
where b	जहां बी

Table 7: Examples of false positive MWNEs.

Tables 5, 6 and 7 depict some of the fully, partially and falsely identified MWNEs respectively. We further investigated the language model voting for the partially identified MWNEs and learnt that the performance can be increased if the language model is trained on a larger but related domain corpora. This algorithm can also easily be extended to other languages, provided the translation model and the transliteration model between the desired language pair and the language models for both the languages are available. Moreover, in the absence of the translation and transliteration models between the desired pair of languages, pivot language based strategy can very easily be incorporated in this approach.

5. Conclusion and Future Work

We are able to suggest a new approach to mine MWNE equivalents from text re-use pair of documents successfully. The approach of jointly estimating MWNEs for a language pair, without having prior knowledge of MWNEs in either of them. The preliminary investigation gives encouraging results for Hindi-English. The algorithm can easily be adapted for the distant language pairs, for which, many cross-language resources are not available directly, but which share a common resource rich pivot language. In future, we intend to evaluate this approach on such language pairs. Though the evaluation is carried on the text re-use document pairs without obfuscation which in nature is noisy parallel text, we believe the algorithm can be extended to the comparable corpora, which we intend to investigate in future.

6. Acknowledgment

The work of the researchers of the Technical University of Valencia has been done in the framework of the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems and it has been partially funded by the European Commission as part of the WIQ-EI IRSES project (grant no. 269180) within the FP 7 Marie Curie People Framework, and by MICINN as part of the Text-Enterprise 2.0 project (TIN2009-13391-C04-03) within the Plan I+D+i.

7. References

Alberto Barrón-Cedeño, Paolo Rosso, Sobha Lalitha Devi, Paul Clough, and Mark Stevenson. 2011. Pan@fire: Overview of the cross-language !ndian text re-use detection competition. In *Working notes of Forum for Information Retrieval Evaluation (FIRE 2011)*, pages 131–139, Mumbai, India, December.

Abhijit Bhole, Goutham Tholpadi, and Raghavendra Udupa. 2011. Mining multi-word named entity equivalents from comparable corpora. In *Proceedings of the 3rd Named Entities Workshop (NEWS 2011)*, pages 65–72, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Parth Gupta and Khushboo Singhal. 2011. Mapping hindi-english text re-use document pairs. In *Working notes of Forum for Information Retrieval Evaluation (FIRE 2011)*, pages 141–146, Mumbai, India, December.

Parth Gupta, Sameer Rao, and Prasenjit Majumder. 2010. External plagiarism detection: N-gram approach using named entity recognizer - lab report for pan at clef 2010. In *CLEF (Notebook Papers/LABs/Workshops)*.

Sarvnaz Karimi, Falk Scholer, and Andrew Turpin. 2011. Machine transliteration survey. *ACM Comput. Surv.*, 43(3):17:1–17:46, April.

Alexandre Klementiev and Dan Roth. 2006. Named entity transliteration and discovery from multilingual comparable corpora. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 82–88, Stroudsburg, PA, USA. Association for Computational Linguistics.

Patrik Lambert and Rafael Banchs. 2005. Data inferred multi-word expressions for statistical machine translation. In *Proc. of Machine Translation Summit X*, pages 396–403, Phuket, Thailand.

Jiahui Liu and Larry Birnbaum. 2008. What do they think?: aggregating local views about news events and topics. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 1021–1022, New York, NY, USA. ACM.

Jong-Hoon Oh, Key-Sun Choi, and Hitoshi Isahara. 2006. A comparison of different machine transliteration models. *J. Artif. Int. Res.*, 27:119–151, October.

Ralf Steinberger, Bruno Pouliquen, and Johan Hagman. 2002. Cross-lingual document similarity calculation using the multilingual thesaurus eurovoc. In *CICLing*, pages 415–424.

Raghavendra Udupa, K. Saravanan, A. Kumaran, and Jagadeesh Jagarlamudi. 2008. Mining named entity transliteration equivalents from comparable corpora. In *CIKM*, pages 1423–1424.

Imed Zitouni and Radu Florian. 2008. Mention detection crossing the language barrier. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 600–609, Stroudsburg, PA, USA. Association for Computational Linguistics.

Projecting Opinion Mining resources across languages and genres

Carlos Rodríguez-Penagos, Jens Grivolla, Joan Codina-Filbá
Barcelona Media Innovació
Av. Diagonal 177, 9th floor, Barcelona 08018, Spain
E-mail: carlos.rodriguez@barcelonamedia.org, jens.grivolla@barcelonamedia.org, joan.codina@barcelonamedia.org

Abstract

Creating language-specific resources to mine opinions in user-generated content can be a laborious task, but even less funded languages have the need for such processing in our increasingly connected world. We describe some experiments in creating Catalan polar lexicons from Spanish resources using automatic word-by-word translation as well as whole corpus Machine Translation for applying bayesian classification methods. Even though some challenges remain in data sparseness and domain adaptation, we believe a practical way of transporting attitude-related contextual information is possible, beyond the more conventional translation of literal lexical meaning.

Keywords: Opinion Mining, lexicon creation, machine translation, Catalan, Spanish

1. Introduction

Sentiment analysis, opinion mining, affective computing, are names for conceptually related technologies which, broadly speaking, aim at detecting and classifying the attitudes and opinions of digital media users. In particular, mining commentaries and interactions expressed online has generated an enormous amount of interest due to the fact that computational exploitation of Social Media and of community-based, data-driven discussions on diverse topics and products has become an important aspect of marketing and business intelligence competencies, since more and more of our activities as citizens, friends and consumers take place in an online environment.

Ubiquity of online collaborative interactions based on the worldwide web also means that user-generated content can have universal reach, and can no longer be confined to local or national contexts. Managing multilingual communication and knowledge is no longer a luxury, especially for multinational organizations and corporations that have a massive client base, and want to be aware, in real time, of what is being said about them and their products in consumer review sites, blogs, forums and twitter microblogging.

Opinion mining (OM) involves at least three specific subtasks: subjectivity identification, polarity detection and intensity estimation (Pang & Lee, 2008). In the case of polarity detection, the objective is to determine the orientation of a given opinionated content. Although two main possible orientations are possible: “positive” and “negative”, it is also common and useful to incorporate the notion of “neutral” opinionated content (Koppel & Schler, 2006). In a simplistic approach, the problem of polarity detection can be approximated as a classification problem, which can be implemented by means of either supervised (Pang et al., 2002; Esuli & Sebastiani, 2006) or unsupervised (Turney, 2002) techniques.

An early and pragmatic approach to detect the semantic orientation of a text is through the use of polar lexicons that list opinion-bearing words and phrases along with their prior polarity. The presence of these attitude cues in a sentence can mean that the opinion expressed has a negative or a positive polarity, although contextual use can determine the final orientation of the phrase, and more sophisticated methods for this computation are needed (Wilson Wiebe & Hoffmann, 2009).

Creating resources for doing opinion mining is a very laborious process, even for the case of accurate polar lexicons limited to a specific domain or genre. Certain customization is needed, since qualifying something as “expensive” can bring positive connotations for some products (“high-quality”), but in some cases it can express a negative attitude towards some other class of products (“low value”). English and some other western languages have had the lion’s share of resources and investment levels needed to develop cutting-edge systems. But in order to avoid a “technological handicap” the gap needs to be filled for less-resourced languages that can’t afford conventional development of these systems, but that definitely have the need for them in a tightly connected world.

In this article we will describe some experiences in using basic machine translation technologies to bootstrap generation of polar lexicons, for a language that lacks them, from resources developed for a closely related one.

2. Semi-automatic generation of OM resources for closely-related languages

Translating literal meaning and translating pragmatic orientation and connotation are very different tasks. Although Machine Learning has advanced considerably with regard to the first aspect, the latter has not received as much attention. The way a language convey attitudes can be nevertheless as much a defining and idiosyncratic trait as, say, a pronominal system.[1]

Spanish and Catalan are closely related romance languages from the Iberian Peninsula that have a unique

[1] Romance languages, for example, have a certain reputation for being more adept to circumlocutions and euphemisms for expressing opinions, but this might only be a topic that is hard to analyze since all languages have these resources and they are not always easy to spot.

and at times uncomfortable cultural symbiosis. The vagaries of history, demography and geography have resulted in the predominance of one at the State level, while at a local level Catalan has slowly regained prominence (and funding) in the last decades. This uneven development exists also (but fortunately not as acutely) with regard to development of Natural Language Processing systems and resources, especially when compared to other "minorized" national languages that can't boast dependency parsing, named-entity recognition or POS tagging at the level that local research groups have achieved for Catalan. Nonetheless, Opinion Mining resources have not been developed yet that match those capabilities. Even for Spanish, the third language in internet with a higher commercial demand, OM-specific resources are scarce, and sometimes are just direct (although consciously adapted) translations of English versions of such resources. A practical way to obtain those resources in a reliable manner was needed, especially since Catalan is one of the most actively used languages online (Moreno, Bel, et. al., 2011).

2.1 Translating and adapting polar lexicons

For our first experiment in creating a Catalan polar lexicon we decided to try word-by-word translation of a manually-created Spanish one, since Catalan and Spanish are lexically very similar. Our small source Spanish lexicon was compiled by a linguist from corpus exploration of customer reviews of banks and mobile phone models. We retained only those entries that were estimated to present a clear polarity independently of the local context of application (but could still be domain dependent). The entries included whole phrases as well as individual words, adjectives, adverbs, verbs, nouns but no specific names or brands. Positive cues included *firme* (firm), *cinco estrellas* (five stars), *ayudar* (help), while negative ones included *inestable* (unstable), *cutre* (tacky), austero (austere), etc.

A total of 530 entries (336 positives and 194 negatives) were compiled and automatically translated to Catalan using a 24,000 entry ES-CA bilingual dictionary that included phrases as well as single words. This approach yielded only 260 direct translations, although conversion of entries into their lemmas when no equivalence was found for the exact word, allowed coverage to increase to 527 unique entries. The low coverage dictionary meant that we could lose some similar entries that could receive the same translation once lemmas were used. But even with a bigger dictionary, the simple fact that each word didn't have a usage context limited the precision of this approach save for highly nonambiguous adjectives such as *limpio* (*net* in Catalan, meaning "clean").

2.2 Using high-throughput MT for data-driven opinion cue classification

Another more promising approach involved using distributional information from a corpus and classifiers to elicit polar lexicons. Opinion mining applications have significantly benefited from the availability of large volumes of annotated data for some languages, even if that data has not been expressly annotated for technology development. For example, in some of the websites where consumer's opinions are collected the users are requested to provide numerical ratings along with their textual inputs. This kind of data constitutes a valuable resource for the implementation of opinion mining application based on supervised learning methods.

Also, machine translation, including both rule-based and statistical approaches, has advanced considerably in the last decade, and is becoming reliable enough for some applications (Ney, 2005). One of the best statistical-based ones for the Catalan-Spanish language pair is the N-II system of the Universitat Politècnica de Catalunya (Marino et.al., 2006). Since no large user-generated texts from consumer review exists for Catalan (to our knowledge), we first automatically translated Spanish language hotel reviews from booking.com, in order to create a parallel collection in our target language. Approximately 61,000 positive and 47,000 negative comments from "good" and "bad" text boxes were translated using the public API for the N-II system, with the consistently good quality of this cutting-edge system helped by what were mostly short and concise comments with a limited number of subjects (prices, room conditions, hotel locations, etc.) and sparse vocabulary. Even though the MT system was not customized for the domain, the comments (e.g., "pricey but well located in the downtown area") were generic enough to be well represented in a statistical approach.

Once extensive and parallel corpus were obtained for Catalan and Spanish, we proceeded to train a naïve Bayes classifier[2] for the two existing categories ("positive" and "negative") in our data. The consistency of the resulting models was tested using train-test data segmentation (75%/25%) and usual metrics. The model allowed calculation of how "informative" or "discriminative" is each used feature with regards to most probable class.

We used this approach iteratively to generate a list of the more discriminative words for positive-negative classification in the comments, by keeping only the 1,000 more discriminative words after the first pass, and then generating a new model that used that restricted word set as possible features, reducing search dimensionality without significantly reducing coverage. These experiments were done both for single words as well as for most discriminative bi-grams (using chi-square, χ^2 distributions to calculate information gain). For Spanish, we repeated the lexicon generation using a version of the booking corpus where original words were substituted for their lemmas, to study how much of a vocabulary reduction could be achieved.

The lists generated showed some intuitively qualificative lexemes, such as "quiet", "cleanliness", etc., but also some other words that were valid only for the specific hotel review domain, such as *formigues* (for "ants" in the

[2] The naive Bayes algorithm assumes "naively" the probabilistic independence of discriminative features when applying the Bayes theorem for classification.

room), *dents* (teeth, for toothbrush in the rooms), etc. To sidestep this effect, we created smaller lexicons that filtered out all entries that did not have a reading as adjectives or adverbs in our dictionaries, reducing our lexicon to between a 35% and 19% of the original total. This step also helped lessen what we call the "Matt Damon-Steve Seagal" effect, that is, the bias found in movie reviews by the mention of good and bad actors that represent positive indicators because of extra-linguistic and encyclopaedic knowledge (and a certain taste in movies, of course).

Tables 1 to 3 present some of the performance metrics on these three parallel corpora training runs, showing positive (pos.) and negative (neg.) instances, training and testing sets, estimated recall (rec.), precision (prec.) and accuracy (acc.) for the classification models, both for unigrams (unig.) and bigrams (big.), as well as the total remaining lexical entries after filtering (filt.) the single-word lexicon by part of speech:

Table 1. BOOKING (SP)

pos.		neg.		train	test
60,921		47,276		81,147	27,050
rec.	prec.	rec.	prec.	acc.	
0.865	0.933	0.92	0.842	0.889	unig.
0.898	0.947	0.935	0.877	0.914	big.
				351 (35%)	filt.

Table 2. BOOKING (SP, lemmas)

pos.		neg.		train	test
60,921		47,276		81,147	27,050
rec.	prec.	rec.	prec.	acc.	
0.869	0.934	0.921	0.845	0.892	unig.
0.894	0.941	0.928	0.872	0.909	big.
				331 (33%)	filt.

Table 3. BOOKING (CA)

pos.		neg.		train	test
60,885		47,183		81,050	27,018
rec.	prec.	rec.	prec.	acc.	
0.856	0.937	0.926	0.833	0.886	unig.
0.89	0.95	0.939	0.868	0.911	big.
				195 (19%)	filt.

3. Previous Work

Bootstrapping language resources from one language into a close one is not a new idea. Carreras, Màrquez, and Padró (2003) suggested generating Named Entity Recognition for Catalan using Spanish resources. For a recent review of multilingual opinion mining efforts and crosslingual bootstrapping, see Banea, Mihalcea and Wiebe (2011). A simple translate-and-adapt approach was taken by Redondo, Fraga *et. al.* (2007) to create a Spanish version of the ANEW sentiment analysis lexicon (Bradley & Lang, 1999).

With regards to using MT on opinionated corpus, Banea, Mihalcea et.al. (2009) generated Romanian and Spanish versions of an annotated English corpus (MPQA) using MT to construct OM resources, using naïve Bayes and Support Vector Machines to evaluate their results, with best precision and recall around 67% for Romanian and 68% for Spanish. A somewhat obvious conclusion they reach is that the quality of the target lexicon is never as high as the one of the source language. Kim and Hovy (2006) used an automatically generated English lexicon to build a German version from a UE translation dictionary via word alignment, with a resulting F-measure for positive polarity of 60% and a somewhat lower one (50%) for negative one.

4. Discussion and Evaluation

To evaluate the performance of the polar lexicons generated with our methodologies, a 200 random sentence sample was obtained for various online text genres where opinion was expressed. For Spanish, these included twitter, blogs, review sites and news, but for Catalan only news documents were obtained. Linguists manually annotated the samples iteratively (until a satisfactory agreement was reached) with five possible categories: positive, negative, mixed (when both polarities could be identified), polar (when a polarity existed, but could not be determined without more context than the available one) or neutral, or non-polar, when no discernible polarity or opinion was recognized.

This gold standard was classified automatically using the polar lexicons with a UIMA pipeline that used the Concept Mapper module, and precision, recall and accuracy were measured against the human standard so that when no polarity was detected neutral category was assigned. For the purposes of this evaluation, the sum of all negative and positive instances constituted the total polar (or opinion-bearing) instances detected, since no distinction between "polar" and "mixed" could be significantly reached using the lexicons.

The results[4] for each language and for different domains are shown in tables 4 through 7, and they show only the results for those texts where a polarity was detected:[5]

Table 4. Catalan lexicons evaluations

	positive		negative	
CA news	**prec.**	**rec.**	**prec.**	**rec.**
Lexicon trans.	42.86	9.38	34.09	57.69
NB words	33.33	72.09	43.75	29.17
NB bigrams	33.33	15.62	9.09	4.35
NB POS filter	62.50	15.15	0.00	0.00

[4] In the tables, *Lexicon trans.* refers to the lexicon translated from Spanish, while: *NB words, NB lemmas, NB bigrams* and *NB POS filter* refer, respectively, to lists of words, word lemmas, bigrams and POS-filtered words derived from naïve Bayes classifiers, in the experiments.

[5] This means, of course, that any subjective-objective imbalance in genres will not readily show up here.

Table 5. Spanish lexicons evaluations (blogs)

	positive		negative	
SP Blogs	**prec.**	**rec.**	**prec.**	**rec.**
NB_booking	24.49	52.17	33.33	66.67
NB bigrams	15.15	23.81	0.00	0.00
NB lemmas	26.32	65.22	30.00	75.00
NB POS filter	36.84	31.82	75.00	60.00

Table 6. Spanish lexicons evaluations (reviews)

	positive		negative	
SP Reviews	**prec.**	**rec.**	**prec.**	**rec.**
NB_booking	76.74	50.00	89.71	58.65
NB bigrams	82.86	43.94	91.67	52.88
NB lemmas	80.49	50.00	85.33	61.54
NB POS filter	92.86	19.70	97.92	45.19

Table 7. Spanish lexicons evaluations (news)

	positive		negative	
SP News	**prec.**	**rec.**	**prec.**	**rec.**
NB_booking	14.89	50.00	10.53	20.00
NB bigrams	30.00	60.00	16.67	11.11
NB lemmas	22.22	58.82	7.14	11.11
NB POS filter	15.79	21.43	28.57	20.00

At first glance it is difficult to discern a clear superiority for any methodology by itself. As is the case in other Information Extraction or classification tasks, with precision and recall there needs to be some kind of compromise, as the better results in one usually result in lower numbers for the other one, and a harmonious mean through F1 calculation could show the better balanced one. Nonetheless, a careful combination of high precision lexicons with ones with good recall would allow better results than simply using single lexicons, even using the better performing ones.

In the Spanish comparisons an expected result is that domain dependency means that best performance, by far, was achieved within the review corpus from which the NB-elicited lexicons were generated. The news domain (unfortunately, the only one we have a gold standard for in Catalan) shows the poorer results, but in a sense this is something we should expect since opinions in a self-proclaimed factual medium are not always overtly expressed, and in any case they are expressed with very different lexical and rhetorical resources than in consumer reviews.

For the Catalan news domain, we were somewhat surprised that our lexicons generated from translated corpora achieved good performance. We speculate that Catalan news media is more opinionated and lexically homogeneous than our Spanish news media representation, but this merits further study. Another notable result is that best performance in the four classes covered was almost evenly distributed among the four lexicons generated (except the bigram ones that work better in the Spanish experiments). At first glance those results suggest that a principled combination of the translated and filtered lexicon with the lemma- and word-based ones generated through Naïve Bayes could achieve good results even in a difficult domain such as the news one. Without a human-verified gold standard for Catalan language customer reviews or blogs, we can only speculate that results would improve in line with what we observe in the Spanish genres, where state-of-the-art precision and recall figures are achieved for customer reviews.

5. Conclusions and Future Work

Our experiences with MT and automatic translation methodologies to transport OM resources between closely related languages have allowed us to test the possibilities both of modern translation techniques and lexical methods for Opinion Mining. Unfortunately, data sparseness for certain languages not only affect direct development of resources, but also evaluation and creation of practical demonstrators with real-world value (as opposed to small-scale laboratory settings). In our case, the lack of available corpus for Catalan opinionated text means not only that we cannot create processing resources directly, but that even when we obtain them through indirect means it is difficult to test them using realistic data sets.

A manual revision of our Catalan lexicons (especially those filtered by POS) shows that most entries truly represent domain-independent linguistic devices for attitude qualification, and would also work reasonably well in more generic domains, although a certain number of domain specific items is also present (e.g., *calefactor*, "warmer"). Overall, polarity assignation coincides with language intuition and competency, for the tested contexts.

We believe these methodologies point in the right direction to generate in a pragmatic and cost-effective manner Opinion Mining resources for languages that are close to one where those resources are more easily available. But some work still needs to be done for domain adaptation,[6] in particular, and for merging and combining different lexicons in order to benefit from the strengths of each one and reduce their corresponding weaknesses. Less-resourced languages that have the same needs as more resourced ones in our exhaustively connected societies can surely benefit from these efforts.

[6] At present we are also experimenting with semantic similarity techniques such as Pointwise Mutual Information measurements from domain-specific corpus to achieve this, but presenting those results here would be off point.

6. Acknowledgements

We are grateful to the UPC team (and in particular to Carlos Henriquez) for allowing use of the N-II system API for these experiments, to Rafael Banchs from the Institute for Infocomm Research (Singapore) for his valuable feedback. The research leading to this publication is financed by ACCESO, and is partially supported by the Spanish Centre for the Development of Industrial Technology under the CENIT program, project CEN-20101037 "Social Media".[7]

7. References

Banea C., R. Mihalcea, and J. Wiebe, "Multilingual Sentiment and Subjectivity Analysis," Multilingual Natural Language Processing, editors Imed Zitouni and Dan Bikel, Prentice Hall, 2011.

Banea C., R. Mihalcea, J. Wiebe, and S. Hassan, "Multilingual subjectivity analysis using machine translation," in Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2008, pp. 127–135.

Bradley M. M. and P. J. Lang (1999), "Affective norms for English words (ANEW): Instruction manual and affective ratings," University of Florida: The Center for Research in Psychophysiology, Gainesville, Florida, USA., Technical report C-1, 1999.

Carreras X., L. Màrquez, and L. Padró (2003), "Named entity recognition for Catalan using Spanish resources," in Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1, 2003, pp. 43–50.

Esuli, A., & Sebastiani, F. (2006). Determining term subjectivity and term orientation for opinion mining. Proceedings of EACL-06, 11th Conference of the European Chapter of the Association for Computational Linguistics. Trento, Italy.

Kim S. M. and E. Hovy (2006), "Identifying and analyzing judgment opinions," in Human Language Technology Conference, NACL 2006, pp. 200–207.

Koppel, M., & Schler, J. (2006). The importance of neutral examples in learning sentiment. Computational Intelligence, 22(2): 100-109

Marino J. B., Banchs R. E., Crego J. M., de Gispert A., Lambert P., Fonollosa J. A. R., andCosta-jussà M. R., "N-gram-based Machine Translation," Computational Linguistics, vol. 32, no. 4, pp. 527–549, 2006.

Mihalcea R., C. Banea, and J. Wiebe (2007), "Learning multilingual subjective language via cross-lingual projections," in NACL proceedings, 2007, vol. 45, p. 976.

Mihalcea R., C. Banea, and J. Wiebe, "Learning multilingual subjective language via cross-lingual projections," in ACL 2007, vol. 45, p. 976.

Moreno A., Bel N., Revilla E., et. al. (2011) Catalan. The META-NET Whitepaper Series on European Languages. http://www.meta-net.eu

Ney, H. (2005). One decade of statistical machine translation: 1996-2005. Proceedings of the 10th MT Summit, Phuket, Thailand..

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2(1-2): 1–135.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, (pp. 79-86).

Redondo J., I. Fraga, I. Padrón, And M. Comesaña (2007), "The Spanish adaptation of ANEW (affective norms for English words)," Behavior Research Methods, vol. 39, no. 3, p. 600, 2007.

Schulz J. M., C. Womser-Hacker, and T. Mandl, "Multilingual Corpus Development for Opinion Mining," in Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, 2010.

Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics, (pp. 417-424). Philadelphia, PA.

Wilson T., J. Wiebe, and P. Hoffmann (2009), "Recognizing Contextual Polarity: an exploration of features for phrase-level sentiment analysis," Computational Linguistics, vol. 35, no. 3, pp. 399–433, 2009.

[7] www.cenitsocialmedia.es

Bologna Translation Service: Constructing Language Resources in the Educational Domain

Arda Tezcan, Joeri Van de Walle and Heidi Depraetere
CrossLang
Woodrow Wilsonplein 7, 9000, Gent, Belgium
{arda.tezcan, joeri.vandewalle, heidi.depraetere}@crosslang.com

Abstract

BTS – Bologna Translation Service – is an ICT PSP EU-funded project which specialises in the automatic translation of study programmes. At the core of the BTS framework are several machine translation (MT) engines through which web-based translation services are offered. Statistical machine translation (SMT) systems form the backbones for all BTS language pairs and for such systems the importance of monolingual and bilingual corpora is undeniable. Unfortunately the lack of readily available domain-specific linguistic resources for various language pairs is one of the major obstacles to build engines with high quality output. In this paper, we report on the ongoing work of language resource construction in the educational domain, focusing on various aspects of this work within the scope of BTS. We also present other relevant BTS components and methods that are used with the aim of exploiting the collected resources and improving MT quality further in the BTS framework.

1. Introduction

There is a continuing and increasing need for educational institutes to provide course syllabi documentation and other educational information in English. Access to translated course syllabi and degree programmes plays a crucial role in the degree to which universities effectively attract foreign students and, more importantly, has an impact on international profiling. Trends show that investment in traditional human translation services is prohibitive, so course materials and degree programmes are often provided in the local language only. The Bologna Translation Service (BTS) aims to provide a solution to this problem by offering a low-cost, web-based, high-quality machine translation (MT) service geared towards this specific use-case. The project will make use of existing rule-based and statistical MT technologies and tailor them in order to produce the best possible quality for syllabus translations. The BTS project will feature the customization, integration and validation of software components and data, and will showcase high-quality MT output for citizens, institutions and businesses, to avail of university programmes of study they are currently unaware of.

Although the primary users of BTS will be universities, the service will also reach students. Users will typically log on to a portal where they can make a request for translating a document or alternatively universities can also submit translation requests via the API. The system can be used purely as an MT service for rapid translation needs. Therefore the post-editing task comes as an optional step in the workflow. Post-editors can either be chosen by the user or be assigned automatically from a pool of post-editors in the BTS system. The manual post-editing step consists of editing and approving translations. The approved translations will be sent to the corresponding TM and back to the user.

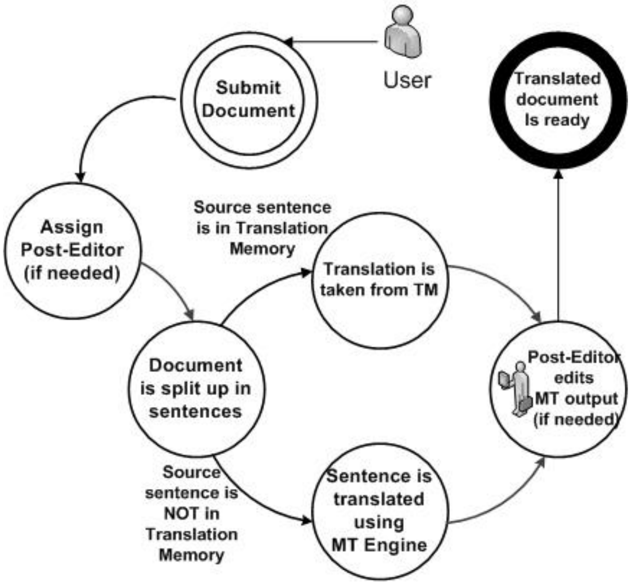

User
Submit Document
Translated document Is ready
Assign Post-Editor (if needed)
Source sentence is in Translation Memory
Translation is taken from TM
Document is split up in sentences
Post-Editor edits MT output (if needed)
Source sentence is NOT in Translation Memory
Sentence is translated using MT Engine

Figure 1: Overview of BTS translation workflow

The consortium responsible for delivering the service consists of a dynamic mix of industry and academia partners, each bringing significant experience and expertise to some facet of the service: CrossLang (Belgium, coordinator), Convertus AB (Sweden), Applied Language Solutions (United Kingdom), Koç University (Turkey), and Eleka Ingeniaritza Linguisitikoa, S.L. (Spain). Communicating on a common platform, each BTS partner specializes on the data collection task, for one or more BTS language pairs.

This paper describes the efforts made by the BTS consortium for constructing language resources in the educational domain for eight language pairs, all of which include English. The other languages that are involved are Finnish, Dutch, German, Portuguese, French, Spanish Turkish and Chinese. The paper focuses on data collection procedures, encountered problems and additional methods that are used in the BTS architecture which aim to improve MT performance by exploiting the full potential of existing language resources or which enable construction of additional resources.

2. SMT Systems in BTS and the Challenge

BTS will provide translation of syllabi and study programmes from seven European languages (Finnish, Dutch, French, Spanish, Portuguese, German and Turkish) to English and from English to Chinese.

Although the fully integrated BTS architecture includes rule-based and statistical MT systems, and makes use of features such as system combination, translation memory (TM), a manual post-editing platform and automatic post-editing, a common backbone for all language pairs is a statistical machine translation (SMT) component. For training and tuning the SMT systems, the freely available Moses toolkit (Koehn, et al., 2007) is used.

The performance of the SMT systems of BTS, like that of all SMT systems, relies primarily on the quantity and the quality of available bilingual data. There are large quantities of such data available in mostly general domain for some of the languages and language pairs in the scope of BTS. However, parallel data directly coming from the educational domain is not readily available for any of the BTS languages and lots of effort must be put into collecting and preparing it so that it can be used for training SMT engines. As a result, data collection emerges as a fundamental task in the project, involving the efforts of all consortium partners.

According to figures provided by “4 International Colleges & Universities” (www.4icu.org), the number of universities, the prospective users for the proposed translation service, amount to over one thousand in the context of our project covering seven countries. Assuming BTS can grow its user base substantially data volumes stored in TMs will grow, giving BTS an opportunity to become a recursively self-improving MT solution in time. This data will not only be useful for translation retrieval but will also be used for obtaining additional corpora. SMT systems can be trained with additional corpora to provide better systems in return to stimulate the use of the service. As a result, the limitations of language resources can be expected to diminish. However, to be able to expand the user base and encourage consistent use of the service, the challenge of the project is to create SMT systems for 8 language pairs for which in-domain data is not available and which provide better translation quality than free alternatives.

3. Data Collection and Corpora Construction

With the aim of providing MT services in a specialised domain, educational documentation in terms of syllabi and study programmes are needed for compiling monolingual and bilingual corpora. Besides the contributions from BTS user group, the consortium also uses web-crawling techniques to extend the collection of in-domain data sets. This work is followed by format handling, code conversions, data clean-up and other preparations to provide ready to use, high-quality language resources.

3.1 Data from User Group

The BTS project started on March 1, 2011 with an initial user group consisting of 8 universities (one university per language) who were willing to make their existing parallel sets of educational documentation available to the project. Meanwhile the user group has expanded to over 50 universities in the first year of the project providing BTS with access to additional user data. The user group members are not only considered as primary contributors for in-domain data but will also be given access to BTS for the entire project duration and will be invited to provide feedback on the usability of the service. So far, collecting the projected amounts of data proved to be a lot more difficult than expected. In their efforts to collect parallel content from the user universities in the educational domain, all partners faced the same kind of difficulties, including:

- Having limited amounts of translated content
- Existence of interpretations or semi-translations rather than literal translations, frequent embedding of non-parallel fragments in multilingual documents
- Problems contacting people with knowledge and access to translated content

3.2 Data via Web-Crawling

A substantial part of the data collection efforts in BTS is also devoted to web crawling. Besides using freely available out-of-domain corpora (e.g. the Europarl corpus (Koehn, 2005), JRC-ACQUIS Multilingual Parallel Corpus (Steinberger, et al., 2006), OPUS corpus (Tiedemann & Nygaard, 2004), etc.), the consortium has also invested heavily in the harvesting of data in the educational domain from the web.

Web-crawling techniques within the scope of BTS mainly focus on detection of candidate websites containing in-domain data and automatic retrieval of candidate content based on URLs. The URL candidates for bilingual content are selected using a technique described in (Almeida et al., 2002) that relies on systematic organization of web pages and is based on pattern recognition using keywords for different languages. For automatic data retrieval, the BTS consortium uses available tools such as HTTrack (Roche, 2007), Wget (Nikšić, 1996) and BootCat Toolkit (Baroni et al., 2004).

Copyright issues can be listed as the major obstacle for web-crawling efforts. During the manual search for relevant data, BTS partners came across various resources with copyright limitations. Thesis abstracts which require permission of individual authors for the use of data can be a given as a good example for copyright limitations on data collection.

On the other hand, to avoid potential copyright issues for automatic retrieval of data, BTS implements the standard opt-out policy which allows publishers of web sites to control what portions (if any) of their site may be crawled, through the use of the Robots Exclusion Protocol (REP)[1].

[1] Reference to REP can be found in the HTML 4.01 specification, Appendix B4.1

As a result, when crawling the web, the crawlers are instructed to abide by the contents of the robots.txt file.

Besides using the English side of the collected parallel data, additional web-crawling efforts have been made to collect monolingual Chinese and additional English data to form the basis of language models (LM) used with the SMT systems in BTS.

3.3 Corpora Construction and Clean-up

After obtaining candidate documents for bilingual (and monolingual) data, corresponding corpora are constructed. To simplify the parallel corpus construction process, the tasks are typically categorized into four areas: Ensuring document level alignment with the use of various pattern recognition methods, format conversions (if necessary), parallel sentence extraction from aligned documents by using freely available and commercial tools (depending on the performance for given language pair(s) and document formats) and cleaning noisy sentence alignments.

Each of the mentioned sub-tasks proves to be challenging in different ways, but the main challenges of corpus construction can be summarized as:

- Working with translated URLs and names for translated documents
- Working with various file formats, with varying complexity in structure and segmentation
- Varying degree of parallelism in the bilingual document pairs
- Varying quality of automatic sentence alignments (different aligners, languages, document types)

Illustrative of some of these difficulties is the data harvesting effort for the ZH—EN language pair. So far, data collection efforts have resulted in approximately 100K good quality aligned sentence pairs, which correspond to approximately only 16% of the total amount of data that was harvested for this language pair.

The corpus construction process results in a set of monolingual and bilingual documents in simple text format (aligned on sentence level). However, the existence of misalignments, interpretations/semi-translations and an arbitrary amount of duplicates in resulting files for the majority of the language pairs of BTS, makes a clean-up procedure an absolute necessity.

Whereas removal of duplicates can be considered a straightforward task, detection and removal of misalignments poses many challenges. For alignment quality analysis the BTS consortium mainly uses a combination of the following heuristics:

- Sentence length comparison: Sentence lengths are not only compared for absolute token size differences but also for the ratio of such differences to the total token size of each sentence.
- Non-word tokens and punctuation analysis: Numbers, special characters and punctuation differences are checked to fine-tune alignment quality.
- MT of source and alignment confidence analysis: Based on the work of Abdul-Rauf and Schwenk (2009), the source segments are translated with the use of an external MT system. Using the automatic evaluation metrics (e.g. METEOR (Banerjee & Lavie, 2005) and edit-distance scores, the MT output is compared to the target side of the parallel corpora.

The acceptance level thresholds are defined manually based on the listed heuristics. As these heuristics are often not enough to draw a clear line between supposedly low and high quality alignments, these threshold values are also confirmed or adjusted by analyzing random subsets of low and high quality alignment candidates. This ensures that an optimal filtering process is in place.

After defining these thresholds the alignments are passed through a filter to eliminate the low quality alignments. As an example, a random subset of 1000 sentences from the final "clean" part of collected NL-EN corpus consists of 98% high quality alignments, whereas the amount of such alignments are observed to be 8% in a random subset of 1000 sentences from the filtered part of this corpus, which is excluded from training of the NL-EN SMT system. An overview of the currently collected and cleaned bilingual and monolingual data can be seen in Table 1 and Table 2 respectively.

Language Direction	**Collected Data (Sentence Pairs)**	**Clean Data (Sentence Pairs)**
DE—EN	91,882	58,399
FR—EN	-[2]	64,060
ES—EN	2,909,157	1,620,358
PT—EN	342,490	212,068
TR—EN	95,314	37,481
FI—EN	531,091	159,809
NL—EN	485,966	223,167
EN—ZH	689,929	112,087

Table 1: Overview of collected and cleaned bilingual data (not final)

Language	**Collected Data (Sentences)**	**Clean Data (Sentences)**
English	4,135,859	2,491,291
Chinese	1,578,819	239,330

Table 2: Overview of collected and cleaned monolingual data (not final)

4. BTS Architecture and Language Specific Applications

As mentioned in Section 2, besides the SMT systems, BTS architecture also consists of other components.

Integrated TMs and the online post-editing platform are common features of BTS for all language pairs. The information stored in the TMs not only allows users to retrieve stored matches but it also allows BTS to obtain

[2] No figures are available for the total amount of collected data for this language pair.

more bilingual and monolingual data to be used in the SMT engines. Similarly, the post-edited data can also be used to boost the quality of the automatic post-editors. Considering the number of potential BTS users and the free availability of the system to user group members for the duration of the project, we expect to obtain additional data for all language pairs within the second year of the project.

Combining freely available out-of-domain data with in-domain data is another method explored by the BTS consortium partners to provide better vocabulary coverage. The applications which have proven to lead to automatic evaluation metric increases include:

- Combining in-domain and out-of-domain data, prior to training translation models, while experimenting with different weights for each data set
- Combining in-domain and out-of-domain phrase tables, during the tuning phase
- Combining in-domain and out-of-domain language models, during tuning phase

Additionally we also observe increases in the automatic evaluation metrics when language-specific language processing methods are used.

Turkish and Finnish are agglutinative languages with complex morphology. Our experiments show that the TR-EN system benefits from full morphological segmentation (Durgar El-Kahlout and Oflazer, 2010). The FI-EN system is another candidate for similar experiments.

Like German, the Dutch language allows arbitrarily long and complex compounds to be formed. With a similar idea of simplifying the morphology, experiments are currently being conducted on the NL-EN system to improve the word coverage and the MT output quality, by using compound splitters. The DE-EN system is another obvious candidate for this application.

5. Conclusions and Future Work

Since BTS offers a translation platform mainly driven by data, data collection and corpora construction tasks lie at the heart of the system.

Building high-quality corpora for a specialised domain from scratch has been a challenge for many and for BTS this challenge is intensified in various ways. The diversity of the languages targeted in BTS, difficulties in accessing user content, difficulties in detecting potential resources on the web, copyright issues, and varying levels of parallelism in obtained parallel documents can be considered as the main challenges in the BTS context.

Data collection is therefore still an ongoing task and is planned to continue for the entire duration of the project with additional strategies applied:

- Expanding web-crawling potential by looking at data from universities outside of Europe
- Expanding data types by including educational data other than syllabi and study programmes.
- Allocating more resources to work on data collection tasks
- Encouraging user group members to post-edit study programmes

The increasing involvement of the user group and the persistent efforts made by the consortium partners together with the range of techniques and methods that are used, let BTS consortium share experience and expertise.

All efforts combined, the BTS consortium is confident that it will be able to construct language resources of sufficient quantity and quality to build high-quality SMT systems.

Although we have succeeded in building competitive SMT systems with the resources constructed from the data collected so far, we hope to further increase the quality of the systems in the future by using the additionally collected data.

6. References

Abdul-Rauf, S. and Schwenk, H. (2009). On the use of comparable corpora to improve SMT performance. In *Proc. of the 12th conference of the European Chapter of the ACL (EACL)*, pages 16-23, Athens, Greece.

Almeida, J. J., Simões, A. M. and Castro, J. A. (2002). Grabbing parallel corpora from the web. *Procesamiento del Lenguaje Natural*, 29:13–20, September.

Banerjee, S., Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization.* University of Michigan, Ann Arbor, MI; pp. 65—72.

Baroni, M., and Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of LREC 2004*. 1313-1316.

Durgar El-Kahlout, I. and Oflazer, K. (2010). Exploiting morphology and local word reordering in english-to-turkish phrase-based statistical machine translation. IEEE Trans. Aud., Sp. and Lang. Proc., 18(6):1313–1322, August.

Koehn et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL, Demonstration Session*.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. *MT Summit 2005.* Phuket, Thailand. pp.79—86.

Niksic, H. (1996). Gnu wget.

Roche, X. (2007). http://www.httrack.com.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., et al. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006).* Genoa, Italy.

Tiedemann, J., and Nygaard, L. (2004). The OPUS corpus - parallel & free. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04).* Lisbon, Portugal.

From scarcity to bounty: how Galateas can turn your scarce short queries into gold

Frédérique Segond, Eduard Barbu, Igor Barsanti, Bogomil Kovachev , Nikolaos Lagos, Marco Trevisan, Ed Vald

Viseo Innovation, Università di Trento, Gonetwork, University of Amsterdam, XRCE, CELI, Bridgeman Art Library
E-mail: fsegond@viseo.net, eduard.barbu@unitn.it, i.barsanti@gonetwork.it ,
B.K.Kovachev@uva.nl,lagos@xrce.xerox.com, trevisan@celi.it, ed.vald@bridgemanart.co.uk

Abstract

With the growth of digital libraries and the digital library federation in addition to partially unstructured collections of documents such as web sites, a large set of vendors are offering engines for retrieving content and metadata via search requests by the end user (queries). In most cases these queries are short unstructured fragments of text in different languages that are difficult to make sense of because of the lack of context. When attempting to perform automatic translation of these queries, using machine learning approaches, the problem becomes worse as aligned corpora are almost inexistent for such types of linguistic data. The GALATEAS European project concentrates on analyzing language-based information from transaction logs and facilitates the development of improved navigation and search technologies for multilingual content access, in order to offer digital content providers an innovative approach to understanding users' behaviour.

Keywords: short query analysis, short query translation, short query understanding

1. Introduction

With the growth of digital libraries and the digital library federation in addition to partially unstructured collections of documents such as web sites, a large set of vendors are offering engines for retrieving content and metadata via search requests by the end user (queries). In most cases these queries are short unstructured fragments of text in different languages that are difficult to make sense of because of the lack of context. When attempting to perform automatic translation of these queries, using machine learning approaches, the problem becomes worse as aligned corpora are almost inexistent for such types of linguistic data.

The GALATEAS European project concentrates on analyzing language-based information from transaction logs and facilitates the development of improved navigation and search technologies for multilingual content access, in order to offer digital content providers an innovative approach to understanding users' behavior. GALATEAS provides two web services based on short query analysis: LangLog and QueryTrans. LangLog focuses on getting meaning out of these lists of queries and is addressed to library/federation/site managers. Unlike mainstream services in this field, GALATEAS services will not consider the standard structured information in web logs (e.g. click rate, visited pages, user's paths inside the document tree) but the *information contained in queries from the point of view of language interpretation*. By subscribing to the LangLog service, federation administrators and managers of content-providing web sites will be able to answer questions such as: "Which are the most commonly searched topics in my collection, according to specific language?"; "how do these topics relate with my catalogue?"; "Which named entities (people, places) are most popular among my users?." QueryTrans has the ambitious and innovative goal of providing the first translation service specially tailored to query translation. The two services are tightly connected: it is only with a successful launch of LangLog that the consortium will gather enough multilingual queries to train the Statistical Machine Translation system adopted by QueryTrans.

While the issue with LangLog is the lack of linguistic resources to analyze short queries, the issue with Query Trans is both the lack of resources, i.e. almost no aligned corpora of short queries, and the tools to translate short queries.

In the next two sections we will give an overview of the resources and methods used to create these two services, highlighting the issues we faced due to lack of resources as well as solutions we adopted.

2. LangLog

LangLog is a system that analyzes and synthesizes the interaction of users with search engines. LangLog illustrates how NLP technologies can be a powerful support tool for market research even when the source of information is a collection of extremely short text documents, each consisting of very few words.

Web systems keep records of their user's interaction. This information is useful in many ways, but its extraction raises many challenges and issues. Facca and Lanzi offer a survey of the topic and show that there are several commercial systems to extract and analyze this information, such as Adobe web analytics, SAS Web Analytics, Infor Epiphany and IBM SPSS. However, none of these contains a linguistic processing component.

Web queries have been the subject of linguistic analysis, to improve the performance of information retrieval systems. For example in [6] the authors experimented with shallow morphological analysis, and in [5] analyzed queries to remove spelling mistakes. LangLog uses raw data describing the interactions of the users with the search engine, such as: the time of the interaction, the text entered by the user, the items picked by the user upon receiving the search results, etc. This data is typically recorded in the log files of the Web server that hosts the search engine. LangLog extracts the data it needs from the log files, provided that they conform to the W3C extended log format, with some additional constraints. In order to develop a prototype of the system, we used Web logs spanning one month of interactions provided by the Bridgeman Art Library. LangLog organizes this data into units called search episodes. Each search episode describes the story of users as they submit a query to the content provider's search engine and picks some (or none) of the search results. We will refer to a picked item as a hit, and we will refer to the text typed by the user as the query. This information alone is valuable to the content provider because it allows discovering which queries were and weren't served with results that satisfied the user.

LangLog analyzes each search episode and records:

1) the language of the query: it may help the content provider decide whether to translate the content into new languages.

2) the lemmas of the query: it is especially important in languages like German and Italian where words have a higher degree of variation. Frequency statistics of keywords help understand what users want, but they are biased towards items associated with words with lesser orthographic and morpho-syntactic variation. For example, three thousand queries for "trousers", two thousand queries for "handbag" and another two thousand queries for "handbags" means that handbags are more popular than trousers, although statistics based on raw words would say otherwise.

3) the named entities referred to in the query: named entities extraction helps the content provider for the same reasons lemmatization does. Named entities are especially important because they identify real-world items that the content provider can relate to, while lemmas do so less often.

4) the category of the query: classification is useful to the content provider because it provides a simpler description of the users' needs. When the target taxonomy is different from the taxonomy used to classify the content provider's products, classification may provide hints as to what kind of needs are not addressed in the catalogue: this is because the classifier classifies the needs expressed by the query, regardless of whether the content provider actually has or does not have items that fulfill those needs. In a similar way, cluster analysis can be used to identify new market segments or new trends in the user's behavior. For example an online book store may discover that one cluster contains many software-related terms, although none of those terms is popular enough to show up in the statistics. If the book store didn't have software-related books, it may decide to acquire some.

In addition to this information, LangLog also performs cluster analysis on search episodes.

3. QueryTrans

QueryTrans is a service specifically targeted to translating queries. Querytrans translates queries posed to the search engine of a content provider into several target languages, without requiring changes to the underlying IR system used and without accessing, at translation time, the content provider's document set. The use of Machine Translation (MT) systems for Cross-Lingual Information Retrieval (CLIR) is widely accepted as one of the best solutions. For instance, Ferro and Peters (2009) show that the best CLIR performance increased from ~55% of the monolingual baseline in 2008 to more than 90% in 2009 for the French and German target languages.

General purpose MT systems are not necessarily adapted for query translation however. This is because statistical MT (SMT) systems trained on a corpus of standard parallel phrases take into account the phrase structure implicitly, while other MT systems tend to use out-of-the-box natural language processing tools originally developed for full phrase analysis. However, the structure of queries is very different from the standard phrase structure: queries are very short and the word order might be different than the typical full phrase query. Take the example of a query like "coupe apollon". While in standard analysis "coupe" would be identified as a verb, in the context of a query it should actually be tagged as a noun, referring to an object. Such a difference may lead to different preprocessing and worse retrieval. In GALATEAS, we adopt two different steps to solving this problem:

1. We adapt a complete and integrated chain of NLP tools to make it suitable for query analysis. The adaptation includes recapitalization, adapted Part of Speech (PoS) tagging, adapted chunking and Named Entities (NEs) recognition. Most of the existing works treat each of these steps independently and address only one of the above issues. In our approach, part of the recapitalization is done during the PoS tagging, in interaction with the NE recognition, which allows us to consider these two steps as interleaved. Moreover, the linguistic processing we propose is generic: corpus-independent (at least most of it, except for NE recognition) and doesn't require access to the document collection (Brun et al. 2012).
2. We adapt SMT model parameters for query translation. To our knowledge, no suitable corpus of parallel queries are available to train an adapted SMT system. Small corpora of parallel queries however can be obtained (e.g. CLEF tracks) or manually created and large volumes of monolingual query data also exist. In our approach the parameters of the SMT models are optimized on the basis of the available query data. This is achieved either directly in the SMT system using the MERT (Minimum Error Rate Training) algorithm, adding monolingual query data to the language model and then optimizing according to the BLEU2 (Papineni et al., 2001) score, or via re-ranking the Nbest translation candidates generated by a baseline system based on new parameters (and possibly new features) that aim to optimize a retrieval metric. It is important to note that both of the proposed approaches allow keeping the MT system independent of the document collection and indexing, and thus suitable for a query translation service.

We evaluated the impact of the first step based on real users' queries, from search logs coming from the Europeana portal. As queries issued on a digital library portal they tend to be very short, referring mostly to artist names, objects, titles, and dates. The impact of the linguistic adaptations is quite significant, in 42% of queries the resulting structure changes. Subsequently, 16% of the query translations are also different. The positive impact of the adapted linguistic processing on the translation quality is evident, for 99 queries the translation is improved when compared to having no linguistic processing. We observe also that 78 queries are better translated after adapting the linguistic processing components.). A lot of the differences are related to the ability to properly identify and handle domain-specific named entities.

The second step was tested on 50 parallel queries from the CLEF AdHoc-TEL2009 task. We have observed that CLIR performance in terms of MAP is improved between 1-2.5 points.

4. Conclusion

In this paper we presented LangLog and QueryTrans, the two linguistics web services (that process short queries in different languages) that have been developed in the GALATEAS project. The quality of the two services strongly depends on the amount of data (short queries in different languages) that can be collected. While in the case of LangLog the issue is the lack of linguistic resources to analyze short queries, in the case of QueryTrans the issue is both the lack of resources, i.e. almost no aligned corpora of short queries, and tools to translate short queries.

We presented in this paper the strategies adopted for both services.

We tested the LangLog system on queries in Bridgeman Art Library. In the future we will test the system on query logs in different domains (e.g. pharmaceutical, hardware and software, etc.) thus increasing the coverage and the significance of the results.

As for QueryTrans we proposed two methods for query-genre adaptation of an SMT model: the first method addressing the translation quality aspect and the second method, the retrieval precision aspect. We believe that the combination of these two methods would be the most beneficial setting, although we were not able to prove this experimentally (due to the lack of training data). In the future we will explore the possibility of combining our adapted SMT model with other state-of-the art CLIR techniques (eg. query expansion with PRF).

5. Acknowledgements

This research was supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430 (Project GALATEAS).

6. References

Salah Aıt-Mokhtar, Jean-Pierre Chanod, and Claude Roux (2002) Robustness beyond shallowness: incremental deep parsing. *Natural Language Engineering*, 8:121–144, June.

Caroline Brun, Vassilina Nikoulina, and Nikolaos Lagos (2012) Linguistically-adapted structural query annotation for digital libraries in the social sciences. In *Proceedings of the 6th EACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Avignon, France, April.

A. Bosca and L. Dini (2010). Language identification strategies for cross language information retrieval. In *CLEF 2010 Working Notes.*

C. Brun and M. Ehrmann. Adaptation of a named entity recognition system for the ester 2 evaluation Campaign (2007). In *IEEE International Conference on Natural Language Processing and Knowledge Engineering*

Nicola Ferro and Carol Peters (2009) CLEF 2009 ad hoc track overview: TEL and persian tasks. In *Working Notes for the CLEF 2009* Workshop, Corfu, Greece.

Facca Federico Michele, Lanzi Pier Luca (2005) Mining interesting knowledge from weblogs: a survey. *Data Knowl. Eng. 53(3): 225-241*

C. Monz and M. de Rijke (2002) Shallow morphological analysis in monolingual information retrieval for dutch, german and italian. *In CLEF 2001.*

K. Papineni, S. Roukos, T. Ward, and W. Zhu. (2001) Bleu: a method for automatic evaluation of machine translation. *IBM Research Report RC22176* (W0109-022).

M. Z. M. Li, Y. Zhang and M. Zhou (2006) Exploring distributional similarity based models for query spelling correction. In *21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL.*

Terminology Extraction, Translation Tools and Comparable Corpora: TTC concept, midterm progress and achieved results

Tatiana Gornostay[a], Anita Gojun[b], Marion Weller[b], Ulrich Heid[b], Emmanuel Morin[c], Beatrice Daille[c], Helena Blancafort[d], Serge Sharoff[e], Claude Méchoulam[f]

Tilde[a], Institute for Natural Language Processing of University of Stuttgart[b]

Laboratoire Informatique de Nantes Atlantique of University of Nantes[c], Syllabs[d], University of Leeds[e], Sogitec[f]

E-mail: scientific-contact@ttc-project.eu

Abstract

The TTC project (Terminology Extraction, Translation Tools and Comparable Corpora) has contributed to leveraging computer-assisted translation tools, machine translation systems and multilingual content (corpora and terminology) management tools by generating bilingual terminologies automatically from comparable corpora in seven EU languages, as well as Russian and Chinese. This paper presents the main concept of TTC, discusses the issue of parallel corpora scarceness and potential of comparable corpora, and briefly describes the TTC terminology extraction workflow. The TTC terminology extraction workflow includes the collection of domain-specific comparable corpora from the web, extraction of monolingual terminology in the two domains of wind energy and mobile technology, and bilingual alignment of extracted terminology. We also present TTC usage scenarios, the way in which the project deals with under-resourced and disconnected languages, and report on the project midterm progress and results achieved during the two years of the project. And finally, we touch upon the problem of under-resourced languages (for example, Latvian) and disconnected languages (for example, Latvian and Russian) covered by the project.

Keywords: language resources, under-resourced languages, disconnected languages, terminology extraction, comparable corpora, computer-assisted translation, machine translation

1. TTC concept and main objectives

The TTC project (Terminology Extraction, Translation Tools and Comparable Corpora) [1] has contributed to leveraging:

- computer-assisted translation (CAT) tools,
- machine translation (MT) systems,
- and multilingual content (corpora and terminology) management tools

by generating bilingual terminologies automatically from comparable corpora in five EU languages belonging to three language families: Germanic (English and German), Romance (French and Spanish), and Baltic (Latvian) as well as outside the European Union: Slavonic (Russian) and Sino-Tibetan (Chinese).

TTC is a three-year project and its main concept is that parallel corpora are scarce resource and comparable corpora can be exploited in the terminology extraction task. The main TTC objectives are as follows:

- to compile and use comparable corpora, for example, harvested from the web;
- to assess approaches that use a minimum of linguistic knowledge for monolingual term candidate extraction from comparable corpora;
- to define and combine different strategies for monolingual term alignment;
- to develop an open web-based platform including solutions to manage comparable corpora and terminology which are also supposed to be available for use in CAT tools and MT systems;
- to demonstrate the operational benefits of the terminology extraction approaches from comparable corpora on CAT tools and MT systems.

2. Parallel vs. comparable corpora

In the end of the 20th century, in natural language processing there was observed a paradigm shift to corpus-based methods exploiting corpora resources (monolingual language corpora and parallel bilingual corpora) with the pioneer researches in bilingual lexicography (for example, Warwick and Russell, 1990) and machine translation (for example, Sadler, 1990).

A parallel corpus is a collection of texts which is translated into one or more languages in addition to the original (EAGLES, 1996). As a rule, parallel corpora are available for certain language pairs, usually including English. This occurs due to the fact that most of natural language processing tools are tailored for English or major European languages (Singh, 2008) in certain domains, for example, the legal domain. The two largest multilingual parallel corpora in the legal domain are:

- the Europarl corpus that covers the language of debates in the European Parliament (Koehn, 2005) and biased to the legal domain;
- the JRC-Aquis corpus that is a huge collection of the European Union legislative documents translated into more than twenty official European languages and includes such rare language combinations as, for example, Estonian-Greek and Maltese-Danish, however still biased to the legal domain (Steinberger et al., 2006).

In view of the quantity of multilingual information that grows exponentially and the need of its translation, parallel corpora can hardly be exploited for facilitating CAT and MT mostly due to their scarceness and limited language

[1] http://www.ttc-project.eu

and domain coverage. This is a well-known and acknowledged fact by the community and it poses a restrictive problem for various translation tasks, be it performed by a human, for example, human and CAT, or a machine and data-driven approaches to MT, for example, statistical machine translation (SMT). Thus one of the main tasks of contemporary natural language processing and corpus linguistics theory and practice is to reduce a linguistic gap between those language pairs that lack cross-language parallel resources and a potential solution to this task is to exploit comparable corpora.

A comparable corpus is a collection of similar texts in more than one language or variety (EAGLES, 1996) and it was introduced to the community in the late 90-ies (Rapp, 1995; Fung, 1995). Since that time, comparable corpora have been actively exploited in different research areas and MT in particular.[2]

The TTC project researches the way in which comparable corpora can be exploited in the terminology extraction task and leveraging translation (CAT and MT) and content (corpora and terminology) management tools.

3. TTC terminology extraction workflow

The TTC multilingual terminology extraction workflow consists of several processing steps (Figure 1).

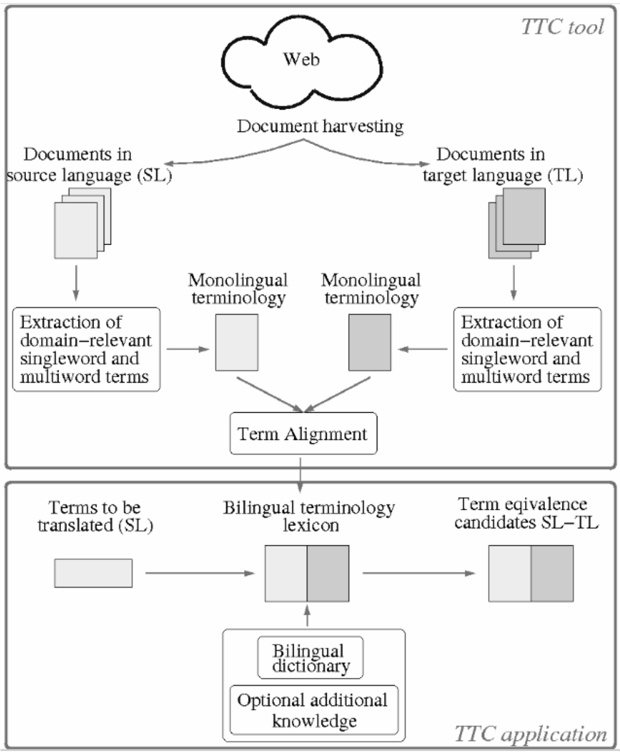

TTC tool
Web
Document harvesting
Documents in source language (SL)
Documents in target language (TL)
Monolingual terminology
Monolingual terminology
Extraction of domain–relevant singleword and multiword terms
Extraction of domain–relevant singleword and multiword terms
Term Alignment
Terms to be translated (SL)
Bilingual terminology lexicon
Term eqivalence candidates SL–TL
Bilingual dictionary
Optional additional knowledge
TTC application

Figure 1. TTC terminology extraction workflow

3.1 Comparable corpora collection

For each TTC language, two domain-specific monolingual corpora have been collected in the wind energy and mobile technology domains.[3] To compile the corpora, we used the focused web crawler developed within the project (Groc, 2011) fed with parallel term seeds in all of the TTC languages. Automatically collected noisy corpora then were manually revised by linguists to get the specialized corpora in the two domains. The size and the quality of the revised corpora vary a lot from language to language. To reach the size of 300 000 running words per domain and per language, the revised corpora were extended with documents manually collected from the web.[4]

To be used in the terminology extraction task, the collected corpora undergo three pre-processing steps:

- tokenization: annotation of word boundaries,
- tagging: annotation of part-of-speech (POS) tags,
- and lemmatization: annotation of lemmas.

3.2 Monolingual terminology extraction

The terminology extraction process in TTC consists of three steps.[5] During the first step, term candidates – single word terms (SWT) and multi-word terms (MWT) – are extracted from the domain-specific corpora collected from the web. The extraction is based on a set of Part-of-Speech patterns (defined for all of the TTC languages) which describe different types of linguistic units, such as nouns (SWT) and adjective + noun, noun + noun, adjective + noun + noun (MWT), etc. During the second step, domain-relevant term candidates are identified. Within the project, we use a frequency-based notion of domain specificity as defined in Ahmad (1992). The final step includes the identification of term variants which may be both synonymous (for example, graphical: *Wind-Energie* ↔ *Windenergie* in German) and related (for example, syntactical: *vēja enerģija* ↔ *vēja un saules enerģija* in Latvian).[6] The output of the extraction component is a list of term candidates sorted descending by their domain specificity values.

3.3 Bilingual terminology alignment

During the next processing step within the TTC terminology extraction workflow, source language and target language monolingual terminologies extracted from comparable corpora are aligned to each other. The result of the alignment step is bilingual domain-specific terminology.

We have proposed to increase the coverage by automatically aligning neoclassical compounds that are extracted from bilingual comparable corpora. Neoclassical

[2] See, for example, the FP7 ACCURAT project research on collecting and using comparable corpora for statistical machine translation (Skadiņa et al., 2012).

[3] TTC comparable corpora are available for download on the website of the University of Nantes under the following link: http://www.lina.univ-nantes.fr/?Linguistic-Resources-from-the.html.

[4] For more information about the TTC domain-specific comparable corpora collected from the web and manually revised, see the project deliverable D2.5 under the following link: http://www.ttc-project.eu/images/stories/TTC_D2.5.pdf.

[5] The project deliverable "D3.4 Set of tools for monolingual term candidate extraction: single and multiword terms and context properties, for example, collocations".

[6] We rely on the set of term variants described in Daille (2005).

compounds are terms that contain at least one neoclassical element (prefix, suffix, and/or root), for example, a term *neuropathy* contains two neoclassical elements *neuro* and *pathy*. For that purpose, a language independent method has been proposed for extracting and aligning neoclassical compounds in two languages. According to this method, neoclassical compounds in the source language are translated compositionally into neoclassical compounds in the target language. For example, the French term *neuropathie* is translated into English by finding the equivalent of each component individually: *neuro → neuro* and *pathie → pathy* and combining these equivalent parts in order to obtain the English translation *neuropathy*. It should be noted, that this translation has to be found in the corpus in order to be considered as correct.

A tool has been developed in order to extract and align neoclassical compounds between two languages from comparable corpora.[7] Experiments were carried out on the following pairs of languages (in both directions): English ↔ French, English ↔ German, and English ↔ Spanish. The results have demonstrated a high precision for all of the translation directions participated in the evaluation. For example, 100 aligned terms were obtained for English↔French with a precision of 98% from the TTC comparable corpora in the wind energy domain.

4. TTC usage scenarios

The resulting bilingual domain-specific terminology can be used as an input to CAT tools and MT systems.[8]

4.1 CAT usage scenario

The extracted bilingual terminology can be integrated into CAT tools which are used by human translators. CAT tools provide the user with target language equivalences and the translator can choose an optimal translation for a source language term. Within the TTC project we evaluate two usage scenarios with CAT involving the English → French language pair in the aeronautic domain and the English → Latvian language pair in the mobile technology domain. The results will be reported by the end of the third year of the project (December 2012).

4.2 MT usage scenario

The output of the TTC term alignment tools can be fed into MT systems as an additional bilingual resource. We explore possibilities of integrating bilingual terminology and domain-specific target language texts (language model data) into statistical machine translation (SMT). First experiments showed that SMT systems using domain-specific texts and bilingual term lists produced by the TTC tools provide better translations than SMT systems without access to these additional knowledge sources (Weller, 2012).

5. TTC & under-resourced languages

One of the TTC languages is Latvian – an under-resourced language of the European Union with approximately 1.5 million native speakers worldwide. For Latvian, the main basic language resources and tools, for example, corpora, lexicons, morphological analysers, etc., are available for processing and evaluation purposes (Skadiņa et al., 2010). More advanced language resources and technologies (for example, discourse corpora, techniques for semantic processing, etc.) are being researched and prototypes are available for some of them.

The resoursefulness of the Latvian language is far from the goal since there is a noticeable gap in language resources and tools of the Latvian language which are a prerequisite of the sustainable development of the language. There are various grammatical characteristics of the Latvian language that make it much more difficult for automatic processing and the two of them (which are most conspicuous and identified as most problematic) are rich inflection and relatively free word order.

Nevertheless, a significant progress has been made in MT for the Latvian language. At the same time, its performance depends on the availability of language resources to a great extent, data-driven approaches in particular. Thus the most researched and developed language pairs in the aspect of SMT are English → Latvian and Latvian → English (Skadiņš et al., 2010). The Latvian-Russian MT is ensured by the rule-based system (Gornostay, 2010).

Nowadays, MT is not anymore considered as a competitor by translators and the task of MT domain adaptation has gained a wide interest. However, for under-resourced languages, the problem of the availability of parallel and even comparable texts still remains an issue. Thus, the Latvian comparable corpus collected within TTC has the smallest size out of the seven TTC languages (cf. 220 823 running words in the Latvian wind energy corpus and 313 954 – in the English wind energy corpus, 314 954 – in the French wind energy corpus, and 358 602 – in the German wind energy corpus). The task of obtaining more corpora for the domain adaptation of the English-Latvian SMT system is currently under consideration within the TTC MT usage scenario.

6. TTC & disconnected languages

Among the so-called “well-researched” language pairs as English-French / German / Latvian / Chinese, French-German / Spanish / Russian and German-Spanish, other TTC working language pairs are Latvian-Russian and Chinese-French which pose the problem of “disconnected languages”. In this situation we deal with two major, or state, languages for which a relatively large amount of monolingual language resources are available but they lack cross-language resources due to their cultural / historical / geographical disconnection.

Despite of the long history of the Latvian and Russian language relationships and their relative similarity

[7] For more information about the Neo-classical MWT detection program for English/French/German, see the project deliverable D4.1 under the link: http://www.ttc-project.eu/images/stories/TTC_D4.1.pdf.

[8] For more information about TTC usage scenarios see Blancafort et al. (2011).

(Gornostay, 2010), there is a considerable lack of Latvian-Russian parallel resources available for research, for example, SMT training and domain adaptation or terminology resource compilation. Within the TTC project, the Latvian-Russian language pair is currently under consideration and the evaluation results of the bilingual terminology extraction for these languages will be reported by the end of June, 2012.

7. Conclusion

TTC is at the beginning of its third year now and so far the project has made significant progress towards the main scientific and technological objectives for the first two years of the project (TTC Annual public report, 2010; 2011).

8. Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 248005.

We would like to thank Inguna Skadiņa for her time and attention to this abstract.

9. References (formatting example)

Ahmad, K. (1992). What is a term? The semi-automatic extraction of terms from text. In *M. Snell-Hornby, F. Poechhacker and K. Kaindl (eds) Translation studies: an interdiscipline*, pp. 267-278.

Daille, B. (2005). Variants and application-oriented terminology engineering. In *Terminology*, Vol. 11, pp. 181-197.

EAGLES (1996). Preliminary recommendations on corpus typology. Electronic resource: http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html.

Fung, P. (1995). A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora. In *Proceedings of the Association for Computational Linguistics*, pp. 236-243.

Gornostay, T. (2010). Latvian-Russian Machine Translation in the System of Social Communication, PhD thesis.

Groc, C. de (2011). Babouk: Focused web crawling for corpus compilation and automatic terminology extraction. In *Proceedings of the IEEE / WIC / ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, Lyon, France, August 2011.

Koehn, P. (2005). Europarl: a parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X*, Phuket, Thailand, 2005.

Rapp, R. (1995). Identifying Word Translations in Non-Parallel Texts. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 320-322.

Sadler, V. and Vendelmans, R. (1990). Pilot implementation of a bilingual knowledge bank. In *Proceedings of Coling-90: Papers presented to the 13th International Conference on Computational Linguistics*, Helsinki, 20-25 August, 1990, Vol. 3, pp. 449-451.

Singh, A.K. (2008). Natural Language Processing for Less Privileged Languages: Where do we come from? Where are we going? In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, Hyderabad, India, January 2008, pp. 7–12.

Skadiņa, I., Aker, A., Glaros, N., Mastropavlos, N., Su, F., Tufis, D., Verlic, M., Vasiljevs, A. and Babych, B. (2012). Collecting and Using Comparable Corpora for Statistical Machine Translation LREC 2012. *In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey.

Skadiņa, I., Auziņa I., Grūzītis N., Levāne-Petrova K., Nešpore G., Skadiņš R., Vasiļjevs A. (2010). Language Resources and Technology for the Humanities in Latvia (2004–2010). In *Proceedings of the Fourth International Conference Baltic (HLT 2010)*, IOS Press, Frontiers in Artificial Intelligence and Applications, Vol. 219, pp. 15-22.

Skadins, R., Goba, K., Sics, V. (2011). Improving SMT with Morphology Knowledge for Baltic Language. In *Proceedings of the Research Workshop of the Israel Science Foundation – Machine Translation and Morphologically-rich Languages*, January 23-27, 2011, Haifa, Israel.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*. Genoa, Italy, 24-26 May 2006, pp. 2142-2147.

TTC Annual public report. (2010). Electronic resource: http://www.ttc-project.eu/images/stories/TTC_Annual_public_report_2010.pdf.

TTC Annual public report. (2011). Electronic resource: http://www.ttc-project.eu/images/stories/TTC_Annual_public_report_2011.pdf.

Warwick, S. and Russell, G. (1990). Bilingual concordancing and bilingual lexicography. In Proceedings of the 4th International Congress EURALEX, Spain, 1990.

Weller, M. (2012). TTC: Terminology Extraction, Translation Tools, and Comparable Corpora. In Proceedings of the 16th Annual Conference of the European Association for Machine Translation, May 28-30, Trento, Italy, 2012. (submitted)

Blancafort, H., Heid, U., Gornostay, T., Mechoulam, C., Daille, B., Sharoff, S. (2011) User-centred Views on Terminology Extraction Tools: Usage Scenarios and Integration into MT and CAT tools. TRALOGY 2012: Translation Careers and Technologies: Convergence Points for the Future”, March 3-4, Paris, France. Electronic resource: http://www.ttc-project.eu/images/stories/TTC_Tralogy_2011.pdf.