Ninth Workshop on Building and Using Comparable Corpora

Workshop Programme

Monday, May 23, 2016

09.15–9.25 ***Opening Remarks***

Session 1: Invited Presentation

09.25–10.30 Ruslan Mitkov
The Name of the Game is Comparable Corpora

10.30–11.00 ***Coffee Break***

Session 2: Building Comparable Corpora

11:00–11:30 Andrey Kutuzov, Mikhail Kopotev, Tatyana Sviridenko and Lyubov Ivanova
Clustering Comparable Corpora of Russian and Ukrainian Academic Texts: Word Embeddings and Semantic Fingerprints

11:30–12:00 Yong Xu and François Yvon
A 2D CRF Model for Sentence Alignment

12:00–12:30 Mehdi Mohammadi
Parallel Document Identification using Zipf's Law

12.30–14.00 ***Lunch Break***

Session 3: Invited Presentation

14.00–15.00 Gregory Grefenstette
Exploring the Richness and Limitations of Web Sources for Comparable Corpus Research

Session 4: Applications of Comparable Corpora

15.00–15.30 Zede Zhu, Xinhua Zeng, Shouguo Zheng, Xiongwei Sun, Shaoqi Wang and Shizhuang Weng
A Mutual Iterative Enhancement Model for Simultaneous Comparable Corpora and Bilingual Lexicons Construction

10:00–10:30 Ana Sabina Uban
Hard Synonymy and Applications in Automatic Detection of Synonyms and Machine Translation

16.00–16.30 ***Coffee Break***

Session 5: Discussion

16:30–17:30 Pierre Zweigenbaum, Serge Sharoff and Reinhard Rapp
Towards Preparation of the Second BUCC Shared Task: Detecting Parallel Sentences in Comparable Corpora

17.30–17.35 ***Closing***

Editors

Reinhard Rapp, University of Mainz, Germany
Pierre Zweigenbaum, LIMSI, CNRS, Université Paris-Saclay, Orsay, France
Serge Sharoff, University of Leeds, UK

Workshop Programme Committee

Ahmet Aker, University of Sheffield (UK)
Hervé Déjean (Xerox Research Centre Europe, Grenoble, France)
Éric Gaussier (Université Joseph Fourier, Grenoble, France)
Vishal Goyal (Punjabi University, Patiala, India)
Gregory Grefenstette (INRIA, Saclay, France)
Silvia Hansen-Schirra (University of Mainz, Germany)
Hitoshi Isahara (Toyohashi University of Technology)
Kyo Kageura (University of Tokyo, Japan)
Philippe Langlais (Université de Montrèal, Canada)
Shervin Malmasi (Harvard Medical School, Boston, MA, USA)
Michael Mohler (Language Computer Corp., US)
Emmanuel Morin (Université de Nantes, France)
Dragos Stefan Munteanu (Language Weaver, Inc., US)
Lene Offersgaard (University of Copenhagen, Denmark)
Ted Pedersen (University of Minnesota, Duluth, US)
Reinhard Rapp (University of Mainz, Germany)
Serge Sharoff (University of Leeds, UK)
Michel Simard (National Research Council Canada)
Pierre Zweigenbaum (LIMSI-CNRS, Orsay, France)

Invited Speakers

Gregory Grefenstette, INRIA Saclay, Université Paris Saclay, France)
Ruslan Mitkov, University of Wolverhampton, UK

Table of Contents

Author Index

Introduction to BUCC 2016

In the language engineering and the linguistics communities, research on comparable corpora has been motivated by two main reasons. In language engineering, on the one hand, it is chiefly motivated by the need to use comparable corpora as training data for statistical Natural Language Processing applications such as statistical machine translation or cross-lingual retrieval. In linguistics, on the other hand, comparable corpora are of interest in themselves by making possible inter-linguistic discoveries and comparisons. It is generally accepted in both communities that comparable corpora are documents in one or several languages that are comparable in content and form in various degrees and dimensions. We believe that the linguistic definitions and observations related to comparable corpora can improve methods to mine such corpora for applications of statistical NLP. As such, it is of great interest to bring together builders and users of such corpora.

Comparable corpora are collections of documents that are comparable in content and form in various degrees and dimensions. This definition includes many types of parallel and non-parallel multilingual corpora, but also sets of monolingual corpora that are used for comparative purposes. Research on comparable corpora is active but used to be scattered among many workshops and conferences. The workshop series on "Building and Using Comparable Corpora" (BUCC) aims at promoting progress in this exciting emerging field by bundling its research, thereby making it more visible and giving it a better platform.

Following the eight previous editions of the workshop which took place in Africa (LREC'08 in Marrakech), America (ACL'11 in Portland), Asia (ACL-IJCNLP'09 in Singapore and ACL-IJCNLP'15 in Beijing), Europe (LREC'10 in Malta, ACL'13 in Sofia, and LREC'14 in Reykjavik) and also on the border between Asia and Europe (LREC'12 in Istanbul), the workshop this year is co-located with LREC'16 in Portorož, Slovenia.

We would like to thank all people who in one way or another helped in making this workshop once again a success. Our special thanks go to Ruslan Mitkov and Gregory Grefenstette for accepting to give invited presentations, to the members of the program committee who did an excellent job in reviewing the submitted papers under strict time constraints, and to the LREC'16 workshop chairs and organizers. Last but not least we would like to thank our authors and the participants of the workshop.

Reinhard Rapp, Pierre Zweigenbaum, Serge Sharoff May 2016

The Name of the Game is Comparable Corpora

Ruslan Mitkov
Research Group in Computational Linguistics
Research Institute in Information and Language Processing
University of Wolverhampton
R.Mitkov@wlv.ac.uk

Abstract

Comparable corpora are the most versatile and valuable resource for multilingual Natural Language Processing. The speaker will argue that comparable corpora can support a wider range of applications than has been demonstrated so far in the state of the art. The talk will present completed and ongoing work conducted by the speaker and colleagues from his research group where comparable corpora are employed for different tasks including but not limited to the identification of cognates and false friends, validation of translation universals, language change and translation of multiword expressions.

Corpora have long been the preferred resource for a number of NLP applications and language users. They offer a reliable alternative to dictionaries and lexicographical resources which may offer only limited coverage. In the case of terminology, for instance, new terms are coined on a daily basis and dictionaries or other lexical resources, however up-to-date they are, cannot keep up with the rate of emergence of new terms. As a result, terminologists (or term extraction programs) seek to analyse the use and/or identify the translation of a specific term using corpora.

Ideally, parallel data would be the best resource both for multilingual NLP applications such as Machine Translation systems and for users such as translators, interpreters or language learners. However, parallel corpora or translation memories may not be available, they may be time-consuming to develop or difficult to acquire as they may be expensive or proprietary. An alternative and more promising approach would be to benefit from comparable corpora which are easier to compile for a specific purpose or task.

Comparable corpora, whether strictly comparable by definition or ‘loosely’ comparable, have already been used in applications such as Machine Translation (Rapp, Sharoff and Zeigenbaum 2016) and term extraction and have been used by translators (Corpas and Seghiri 2009). The good news is that comparable corpora can facilitate almost any multilingual application and can beneficial to almost any language user. The view of the speaker is that comparable corpora are the most versatile, valuable and practical resource for multilingual NLP. The invited talk at the BUCC workshop at LREC’2016 will show that comparable corpora can offer more in terms of value and can support a wider range of applications than has been demonstrated so far in the state of the art. The talk will present completed and ongoing work conducted by the speaker and his colleagues at the Research Group in Computational Linguistics at the University of Wolverhampton in the domain of comparable corpora.

The talk will start with a discussion of the notion of comparable corpora and issues related to their use and compilation, and will briefly outline work by the speaker and his colleagues on the methodology related to the extraction of comparable documents and the building of purpose-specific comparable corpora.

Next the work carried out by the author on the automatic identification of cognates and false friends using comparable data will be presented. This will be followed by the presentation of three novel approaches developed by the speaker which use comparable data but do not resort to any dictionaries or parallel corpora, together with extensive evaluations of their performance.

The speaker will then focus on the use of purpose-built comparable corpora and NLP methodology in a project whose objective was to test the validity of so-called translation universals. In particular, the experiments on validating the universals of simplification, convergence and transfer will be detailed.

Following from this study, the speaker will outline the work on the use of comparable corpora to track language change over time, in particular the recent changes in lexical density and lexical richness in two consecutive thirty-year time periods in British English (1931–1961 and 1961–1991) and in American English from the 1960s to the 1990s (1961–1992).

Finally, the speaker will share the latest results from his work with colleagues on the use of comparable corpora for extracting and translating multiword expressions. The methodology developed does not rely on any dictionaries or parallel corpora, nor does it use any (bilingual) grammars. The only information comes from comparable corpora, inexpensively compiled with the help of the ACCURAT toolkit (Su and Babych 2012a) where only documents above a specific threshold were considered for inclusion. The presentation will conclude with the results of an interesting experiment as part of this study which sought to establish whether large loosely comparable data would yield better results than smaller but strictly comparable corpora.

Bibliographical References

Corpas, G. 2008. Investigar con corpus en traducción: los retos de un nuevo paradigma. Frankfurt: Peter Lang.

Corpas, G. and Seghiri M. 2009. "Virtual Corpora as Documentation Resources: Translating Travel Insurance Documents (English-Spanish)". In Beeby, A., Sánchez, P. and Rodríguez P. (Eds) Corpus Use and Learning to Translate. Proceedings from the CULT Conference, Barcelona, Spain, John Benjamins, 75-107.

Corpas, G., Mitkov R., Afzal, N. and Garcia Moya, L.

2008. "Translation universals: do they exist? A corpus-based and NLP approach to convergence". Proceedings of the LREC’2008 Workshop on Building and Using Comparable Corpora.

Corpas, G., Mitkov R., Afzal, N. and Pekar, V. 2008. "Translation universals: do they exist? A corpus-based NLP study of convergence and simplification". Proceedings of the AMTA’2008 conference, Honolulu, Hawaii, 75-81.

Costa, H., Corpas, G., Mitkov, R. and M. Seghiri. 2015. "Towards a Web-based Tool to Semi-automatically Compile, Manage and Explore Comparable and Parallel Corpora". In Proceedings of the 7th International Conference of the Iberian Association of Translation and Interpreting Studies (AIETI’2015). Malaga, Spain

Costa, H., Corpas, G. and R. Mitkov. 2015. "Measuring relatedness between documents in comparable corpora". In Proceedings of the 11th International Conference on Terminology and Artificial Intelligence (TIA'15), Granada, Spain, 29-37.

Fung, P. and Cheung, P. 2004. "Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus". In Proceedings of the 20th international conference on Computational Linguistics (COLING), Geneva, Switzerland

Ilisei, I., Inkpen, D., Corpas, G., and Mitkov, R. 2012. "Romanian Translational Corpora: Building Comparable Corpora for Translation Studies". In Proceedings of the 5th Workshop on Building and Using Comparable Corpora (5th BUCC), held in conjunction with LREC 2012, Istanbul, Turkey, 56-61.

Kilgarriff , A. 2010. "Comparable Corpora Within and Across Languages, Word Frequency Lists and the KELLY Project". In Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC’2010, Malta.

Mendoza Rivera, O., Mitkov R. and G. Corpas Pastor. 2013. "A Flexible Framework for Collocation Retrieval and Translation from Parallel and Comparable Corpora" In Proceedings of the International Workshop on Multiword units in Machine Translation and Translation Technology. Nice, France.

Mitkov R., Pekar V., Blagoev D. and Mulloni A. 2008. "Methods for extracting and classifying pairs of cognates and false friends ". Machine Translation. 21 (1), 29-53.

Mitkov, R. 2016. "The benefit of comparable corpora: automatic translation of multiword expressions without translation resources" (forthcoming). In Corpas, G. and Seghiri, M. (Eds). Corpus-based approaches to translation and interpreting: from theory to applications. Peter Lang

Pekar V., Mitkov R., Blagoev D. and Mulloni A. 2008. "Finding Translations for Low-Frequency Words in Comparable Corpora". Machine Translation, 20 (4), 247-266.

Pekar V., Mitkov R., Blagoev D. and Mulloni A. 2007. "Finding Translations for Low-Frequency Words in Comparable Corpora. " Proceedings of the CONTEXT-07 Workshop on "Contextual Information in Semantic Space Models" (CoSmo-2007), Roskille, Denmark, 17-25.

Pinnis, M., Ion, R., Ştefănescu, D., Su, F., Skadiņa, I., Vasiļjevs, A., and Babych, B. 2012. "ACCURAT Toolkit for Multi-Level Alignment and Information Extraction from Comparable Corpora. Proceedings of the ACL 2012 System Demonstrations, Jeju, Korea, 91-96.

Rapp, R, Sharoff, S. and Zweigenbaum, P. (Eds). 2016. Special Issue on using comparable corpora for Machine Translation. Journal of Natural Language Engineering, 22(4). (forthcoming).

Skadiņa, I., Aker, A., Mastropavlos, N., Su, F., Tufiş, D., Verlic, M., Vasiļjevs, A., Babych, B., Clough, P., Gaizauskas, R., Glaros, N., Paramita, M. and Pinnis, M. 2012. "Collecting and Using Comparable Corpora for Statistical Machine Translation". Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC’12), Istanbul, Turkey, 438–445.

Štajner, S and Mitkov, R. 2012. Using Comparable Corpora to Track Diachronic and Synchronic Changes in Lexical Density and Lexical Richness, in Proceedings of the 5th Workshop on Building and Using Comparable Corpora (5th BUCC), held in conjunction with LREC 2012, Istanbul, Turkey, 88-97.

Stambolieva, E. 2012. Compiling Comparable Corpora: A Machine Learning Approach. MSc Dissertation, University of Wolverhampton.

Su, F. and Babych, B. 2012a. "Measuring Comparability of Documents in Non-Parallel Corpora for Efficient Extraction of (Semi-)Parallel Translation Equivalents". Proceedings of the EACL'12 Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra), Avignon, France, 10-19.

Su, F. and Babych, B. 2012b. "Development and Application of a Cross-language Document Comparability Metric". Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey. 3956-3962.

Taslimipoor, S., Mitkov, R., Corpas Pastor, G. and Fazly, A. 2016. "Bilingual Contexts from Comparable Corpora to Mine for Translations of Collocations." In A. Gelbukh (Ed.): CICLing 2016, LNCS vol. 9623. Springer, Heidelberg.

Yapomo, M., Corpas. G., Mitkov, R. 2012. "CLIR- and ontology-based approach for bilingual extraction of comparable documents Proceedings of the 5th Workshop on Building and Using Comparable Corpora, LREC 2012, Istanbul, Turkey, 121-125.

Clustering Comparable Corpora of Russian and Ukrainian Academic Texts: Word Embeddings and Semantic Fingerprints

Andrey Kutuzov, Mikhail Kopotev, Tatyana Sviridenko, Lyubov Ivanova
University of Oslo, University of Helsinki, National Research University Higher School of Economics
Oslo, Helsinki, Moscow
andreku@ifi.uio.no, mihail.kopotev@helsinki.fi, sviridenkot@gmail.com, luben92@gmail.com

Abstract

We present our experience in applying distributional semantics (neural word embeddings) to the problem of representing and clustering documents in a bilingual comparable corpus. Our data is a collection of Russian and Ukrainian academic texts, for which topics are their academic fields. In order to build language-independent semantic representations of these documents, we train neural distributional models on monolingual corpora and learn the optimal linear transformation of vectors from one language to another. The resulting vectors are then used to produce 'semantic fingerprints' of documents, serving as input to a clustering algorithm. The presented method is compared to several baselines including 'orthographic translation' with Levenshtein edit distance and outperforms them by a large margin. We also show that language-independent 'semantic fingerprints' are superior to multi-lingual clustering algorithms proposed in the previous work, at the same time requiring less linguistic resources.

Keywords: word embeddings, text clustering, comparable corpora, academic texts, cross-lingual transformations

1. Introduction

This research addresses the problem of representing the semantics of text documents in multi-lingual comparable corpora. We present a new approach to this problem, based on neural embeddings, and test it on the task of clustering texts into meaningful classes depending on their topics. The setting is unsupervised, meaning that one either does not have enough annotated data to train a supervised classifier or does not want to be limited with a pre-defined set of classes. There is a lot of sufficiently good approaches to this problem in the case of mono-lingual text collections, but the presence of multiple languages introduces complications.

When a text collection contains documents in several languages, it becomes impractical to simply represent the documents as vectors of words occurring in them ("bag-of-words"), as the words surface forms are different, even in closely-related languages. Thus, one has to invent means to cross the inter-lingual gap and bring all documents to some sort of shared representation, without losing information about their topics or categories.

Of course, one obvious way to solve this problem is to translate all documents into one language, and then apply any clustering algorithm. However, this requires either buying human/machine translation services (which can be expensive if you deal with large text collection) or training own statistical machine translation model (which as a rule requires big parallel corpus). This is the reason to search for other solutions.

In this paper, a novel way of reducing the problem of cross-lingual document representation to a monolingual setting is proposed. Essentially, we train Continuous Bag-of-Words models (Mikolov et al., 2013b) on large comparable monolingual corpora for two languages our dataset consists of. This provides us with vector representations of words, allowing to measure their semantic similarity. Then, a linear transformation matrix from vectors of language *A* to vectors of language *B* is learned, using a small bilingual dictionary as training data. This matrix is then employed to 'project' word and document representations from semantic space of language *A* to semantic space of language *B*. It allows not only quite accurate 'translation' of words, but also of document '*semantic fingerprints*' (dense representations of document semantics, calculated as an average of the trained distributional vectors for all the words in document).

This approach is evaluated in a setting, where the input is a collection of documents in several languages and some number of topics to which these documents belong (we also have large monolingual corpora to train distributional models on). For each document, we are given its language, but not its topic. The task is to cluster this collection so that documents belonging to one topic were clustered together, independent of their language. Note that we are interested in clustering the collection as a whole, not each language separately (which is trivial).

Our evaluation data consists of comparable corpora of Russian and Ukrainian academic texts. On this material, we show that the '*translated semantic fingerprints*' method represents documents in different languages precisely enough to allow almost exact clustering according to document topics, with only 5% of incorrect assignments. It significantly outperforms both naive bag-of-words baseline and the not-so-naive method of 'orthographic translation' based on Damerau-Levenshtein distance, even enriched with dictionary mappings. At the same time, it does not require large parallel corpora or a ready-made statistical machine translation model.

The rest of the paper is structured as follows. In Section 2. we describe the foundations of our approach and the related work. Section 3. introduces the employed corpora and the story behind them. Section 4. is dedicated to learning the transformation matrix, and Section 5. describes our experimental setting and evaluation results. We discuss the findings in Section 6. and conclude in Section 7., also suggesting directions for future work.

2. Related Work

Clustering multi-lingual documents has received much attention in natural language processing. Among approaches not using some form of machine translation, one can mention (Mathieu et al., 2004), who essentially employ a bilingual dictionary to bring some words in the documents to a language-independent form and then to perform clustering. In the section 5. we show that our approach based on neural embeddings significantly outperforms their reported results.

(Wolf et al., 2014) proposed training joint multi-lingual neural embedding models. Theoretically, this can be used to achieve our aim of language-independent semantic representations for documents. Unfortunately, it demands a large word-aligned parallel corpus. This is not the case with the more recent *Trans-gram* approach introduced in (Coulmance et al., 2016), also able to learn multi-lingual models. However, it still needs sentence-aligned corpora to train on (in the size of millions of paired sentences). Large parallel corpora (whether word- or sentence-aligned) are often a scarce resource, especially in the case of under-represented languages.

The approach described in this paper takes as an input only comparable monolingual corpora and bilingual dictionaries in the size of several thousand word pairs. Such resources are much easier to find and evaluate. We employ the idea of learning a linear transformation matrix to map or project word embeddings from the semantic space of one language to that of another. This idea was first proposed in (Mikolov et al., 2013a), who applied it to lexical translation between English, Spanish, Czech and Vietnamese. We extend it from continuous representations of single words or collocations to '*semantic fingerprints*' of documents as a whole.

3. Academic texts as Comparable Corpora

The Russian and Ukrainian languages are mainly spoken in Russian Federation and the Ukraine and belong to the East-Slavic group of the Indo-European language family. They share many common morphosyntactic features: both are SVO languages with free word order and rich morphology, both use the Cyrillic alphabet and share many common cognates.

Both Russia and the Ukraine have common academic tradition that makes it easier to collect corpora, which are comparable in terms of both genre and strictly defined academic fields. We work with such a corpus of Russian and Ukrainian academic texts, initially collected for the purposes of cross-lingual plagiarism detection. This data is available online through a number of library services, but unfortunately cannot be republished due to copyright limitations.

The Ukrainian subcorpus contains about 60 thousand extended summaries (Russian and Ukrainian 'автореферат', '*avtoreferat*') of theses submitted between 1998 and 2011. The Russian subcorpus is smaller in the number of documents (about 16 thousand, approximately the same time period), but the documents are full texts of theses, thus the total volume of the Russian subcorpus is notably larger: 830 million tokens versus 250 million tokens in the Ukrainian one. Generally, the texts belong to one genre that can be defined as post-Soviet expository academic prose, submitted for academic degree award process.

The documents were converted to plain text files from MS Word format in the case of the Ukrainian subcorpus and mainly from OCRed PDF files in the case of the Russian subcorpus. Because of this, the Russian documents often suffer from OCR artifacts, such as words split with line breaks, incorrectly recognized characters and so on. However, it does not influence the resulting model much, as we show below.

Both Ukrainian and Russian documents come with meta data allowing to separate them into academic fields, with economics, medicine and law being most frequent topics for the Ukrainian data and economics, history and pedagogy dominating the Russian data.

For evaluation, 3 topics were used, distant enough from each other and abundantly presented in both subcorpora: economics, law and history. We randomly selected 100 texts in each language for each topic. As an average length of Russian texts is significantly higher (them being full theses), we cropped them, leaving only the first 5 thousand words, to mimic the size of the Ukrainian summaries. These 600 documents in 3 classes are used as a test set (see Section 5. for the description of the conducted experiments).

The corpora (including test set) were PoS-tagged[1]. Each word was replaced with its lemma followed by a PoS-tag ('диссертация_S', 'диссертація_N'). Functional parts of speech (conjunctions, pronouns, prepositions, etc.) and numerals were removed from the texts.

4. Learning to Translate: Ukrainian-to-Russian transformations

As already stated, our main proposal is using neural embedding models to 'project' documents in one language into the semantic space of another language. For this, we first trained a Continuous Bag-of-Words (CBOW) and a Continuous SkipGram model (Mikolov et al., 2013b) for each of our monolingual subcorpora. The models were trained with identical hyperparameters: vector size of 300 components[2], symmetric window of 2 words, negative sampling with 10 samples, 5 iterations over the corpus, no down-sampling. The only language-dependent difference was that for the Ukrainian model we ignored words with the corpus frequency less than 10 and for the Russian model this threshold was set to 15 (as the Russian corpus is 3 times larger). All in all, the final Ukrainian model recognizes 429 215 words and the Russian one 271 720 words. Training was performed using CBOW and SkipGram implementation in *Gensim* library (Řehůřek and Sojka, 2010).

After the models were trained, we followed the path outlined in (Mikolov et al., 2013a) to learn a linear transformation matrix from Ukrainian to Russian. First, we extracted

[1] We used *Mystem* (Segalovich, 2003) for Russian and *Ugtag* (Kotsyba et al., 2009) for Ukrainian.

[2] (Mikolov et al., 2013a) suggest to use larger vector size for the source language model; however, we leave it for the future work.

all noun pairs from Russian-Ukrainian bilingual dictionary (Ganich and Oleynik, 1990), with the constraint that their frequency in our corpora was above the already mentioned thresholds 15 and 10 for Russian and Ukrainian words correspondingly. That made it a list of about 5 thousand pairs of nouns being translations of each other.

For all these words, their vectors were found in the models corresponding to the words' languages. It provided us with a matrix of 5 thousand of 300-dimensional Ukrainian vectors and the matrix of corresponding 5 thousand of 300-dimensional Russian vectors. This data served as a training set to learn an optimal transformation matrix. The latter is actually a 300x301 matrix of coefficients, such that when the initial Ukrainian matrix is multiplied by this transformation matrix, the result is maximally close to the corresponding Russian matrix. This transformation matrix has 301 (not 300) columns, because we add one component equal to 1 to each vector, as a bias term.

Producing the transformation matrix is a linear regression problem: the input is 301 components of Ukrainian vectors (including the bias term) and the output is 300 components of Russian vectors. As we need 300 values as an output, there are actually 300 linear regression problems and that's why the resulting matrix size is 300x301 (301 weights for each of 300 components).

There are two main ways to solve a linear regression problem: one can either learn the optimal weights in an iterative way using some variant of gradient descent, or one can solve it numerically without iteration, using normal equation. For English and Spanish, (Mikolov et al., 2013a) used stochastic gradient descent. However, normal equation is actually less error-prone and is guaranteed to find the global optimum. Its only disadvantage is that it becomes very computationally expensive when the number of features is large (thousands and more). However, in our case the number of features is only 301, so computational complexity is not an issue.

Thus, we use normal equation to find the optimal transformation matrix. The algebraic solution to each of 300 normal equations (one for each vector component i) is shown in the Equation 1:

$$\boldsymbol{\beta}_i = (\mathbf{X}^\intercal * \mathbf{X})^{-1} * \mathbf{X}^\intercal * y_i \quad (1)$$

where $\mathbf{X}$ is the matrix of 5 thousand Ukrainian word vectors (input), y_i is the vector of the ith components of 5 thousand corresponding Russian words (correct predictions), and $\boldsymbol{\beta}_i$ is our aim: the vector of 301 optimal coefficients which transform the Ukrainian vectors into the ith component of the Russian vectors.

After solving such normal equations for all the 300 components i, we have the 300x301 linear transformation matrix which fits the data best.

This matrix basically maps the Ukrainian vectors into the Russian ones. It is based on the assumption that the relations between semantic concepts in different languages are in fact very similar (*students* are close to *teachers*, while *pirates* are close to *corsairs*, and so on). In continuous distributional models which strive to represent these semantic spaces, mutual 'geometrical' relations between vectors representing particular words are also similar across models (if they are trained on comparable corpora), but the exact vectors for words denoting one and the same notion are different. This is because the models themselves are stochastic and the particular values of vectors (unlike their positions in relation to each other) depend a lot on technical factors, including the random seed used to initialize vectors prior to training. In order to migrate from a model **A** to another model **B**, one has to 'rotate and scale' **A** vectors in a uniform linear way. To learn the optimal transformation matrix means to find out the exact directions of rotating and scaling, which minimize prediction errors.

Linguistically speaking, once we learned the transformation matrix, we can predict what a Russian vector would most probably be, given a Ukrainian one. This essentially means we are able to 'translate' Ukrainian words into Russian, by calculating the word in the Russian model with the vector closest to the predicted one.

We had to choose between CBOW or Continuous Skip-Gram models to use when learning the transformation matrix. Also, there was a question of whether to employ regularized or standard normal equations. Regularization is an attempt to avoid over-fitting by trying to somehow decrease the values of learned weights. The regularized normal equation is shown in 2:

$$\boldsymbol{\beta}_i = (\mathbf{X}^\intercal * \mathbf{X} + \lambda * L)^{-1} * \mathbf{X}^\intercal * y_i \quad (2)$$

Comparing to 1, it adds the term $\lambda * L$, where L is the identity matrix of the size equal to the number of features, with 0 at the top left cell, and λ is a real number used to tune the influence of regularization term (if $\lambda = 0$, there is no regularization).

To test all the possible combinations of parameters, we divided the bilingual dictionary into 4500 noun pairs used as a training set and 500 noun pairs used as a test set. We then learned transformation matrices on the training set using both training algorithms (CBOW and SkipGram) and several values of regularization λ from 0 to 5, with a step of 0.5. The resulting matrices were applied to the Ukrainian vectors from the test set and the corresponding Russian 'translations' were calculated. The ratio of correct 'translations' (matches) was used as an evaluation measure. It came out that regularization only worsened the results for both algorithms, so in the Table 1 we report the results without regularization.

For reference, we also report the accuracy of 'quazi-translation' via Damerau-Levenshtein edit distance (Damerau, 1964), as a sort of a baseline. As already stated, the two languages share many cognates, and a lot of Ukrainian words can be orthographically transformed into their Russian translations (and vice versa) by one or two character replacements. Thus, we extracted 50,000 most frequent nouns from our Russian corpora; then for each Ukrainian noun in the bilingual dictionary we found the closest Russian noun (or 5 closest nouns for @5 metric) by edit distance and calculated how often it turned out to be the correct translation. As the Table 1 shows, notwithstanding the orthographic similarity of the two languages, CBOW consistently outperforms this approach even on the test set. On the training set, its superiority is even more obvious.

Table 1: Translation accuracy

	CBOW		**SkipGram**		**Edit distance**
	Training	Test	Training	Test	
@1	0.648	**0.57**	0.545	0.374	0.549
@5	0.764	**0.658**	0.644	0.486	0.619

As for comparison between learning algorithms for matrix translation, CBOW-based transformation matrix is again the winner, with 57% matches on the test set and 65% matches on the training set, beating SkipGram in both. Note that in the context of this task, SkipGram models seem to have problems with actually learning the optimal transformation matrix for unseen data: on the test set they perform even worse than the edit distance approach.

CBOW is also consistently better if we consider cases when the correct word is among 5 nearest neighbors of the predicted vector to be matches as well (accuracy @5). This is an important metrics, because quite often the 'translation' is not exactly the corresponding word from the dictionary, but still a very semantically similar one, while the dictionary translation is the second or the third by its cosine similarity to the predicted vector. It means that in fact the 'semantic translation' is successful, as the concept is correct. For example, our algorithm translates the Ukrainian noun 'гетьман' *hetman* into Russian 'царь' *tzar*, while the correct translation 'гетман' is the second nearest neighbor.

Notwithstanding the fact that the transformation matrix was trained exclusively on nouns, it correctly 'translates' adjectives and verbs as well (we did not experiment with other parts of speech though). However, it tends to 'substantivize' them: for example, the Ukrainian verb 'розробити' *to develop* is transformed into a Russian vector, which is closer to the noun 'разработка' *development* than to the corresponding verb.

Thus, at least main parts of speech seem to share a common Ukrainian-to-Russian projection matrix, supporting the view that semantic spaces for different languages are in comparatively simple linear relations to each other. In the following clustering experiments we employed CBOW-based transformation matrix and consequently CBOW models for Russian and Ukrainian.

We also applied the same transformation matrix to the document-level 'semantic fingerprints'. These fingerprints are simple average vectors of all words that the document contains. Thus, if our models have vector size 300, the resulting fingerprints are 300-dimensional vectors as well. These vectors can be transformed with the same matrix. As we show in the Section 5., the cross-lingual linear relations hold not only for words, but for these semantic fingerprints as well.

5. Experiment Design and Evaluation

We evaluate the cross-lingual representations described above on the task of clustering a set of documents. Recall that our test set consists of randomly selected 600 documents, equally divided between the topics of economics, law and history, and Russian and Ukrainian languages. Thus, we have 100 Ukrainian law texts, 100 Russian law texts, etc. The average length of the texts is 4000 word tokens.

We aim to find such a representation for documents which would reveal their topical structure independent of the language. It can be tested by clustering the whole collection in an unsupervised way into 3 clusters (in our setting, the number of topics is a given parameter), and finding out to what extent these clusters correspond to the topical classes: law, economics and history. This correspondence can be calculated by mapping the resulting clusters into topics judging by where the majority of documents belonging to this or that topic were assigned. For example, if more than 100 history documents were assigned to the cluster 0, we map this cluster to the history topic, etc. Then, the ratio of incorrect assignments is calculated, as percentage from the total number of documents. This is our primary evaluation measure. All the clustering experiments below are performed using a well-established *K-means* algorithm (Hartigan, 1975) with Euclidean distances. We intentionally employ the most basic clustering algorithm to make the difference of the underlying representations more visible.

The lemmatized documents were represented as bags-of-words. To reduce the dimensionality of such representations and to filter out unimportant noise words, some sort of feature selection is often used. We employed the most basic variant of it: frequency threshold, where the words are ranked by their frequencies in the whole document collection, and only top x are then used in constructing vector representations. We empirically chose $x = 500$, as several values from 100 to 1000 which we tried (with the step of 100) resulted in worse performance, independent of the approaches tested. Note that the sets of 500 most frequent words were selected for each topic separately, to avoid the situation when some topics are under-represented, because words related to them are not frequent. Then the union of these sets was used as the final vocabulary (resulting in vectors of about 800...900 dimensions, depending on the particular method used). Initially, binary vectors were constructed (a word is either present in the document or not), but we also tested count vectors, which store words' per-document frequencies; see below.

In order to make sure that the topical division is indeed manifested in the documents, we first clustered Ukrainian and Russian corpora separately, using the binary bag-of-words representations described above. This gave only 4.7% incorrect assignments for the Ukrainian texts and 34.7% incorrect assignments for the Russian part of the test set. Thus, for Ukrainian the division is almost perfect, while for Russian it is manifested less clearly (it seems that economics and law are consistently mixed up), but still the overwhelming majority of documents is clustered according to the topics. It means that the test set does contain information to correctly cluster the documents on a monolingual level, and it makes sense to try to achieve comparable (or at least not much worse) results in the cross-lingual experiments. Note that one can't simply cluster the documents in Russian and in Ukrainian separately to achieve our aim: even if the clusterings are ideal, there will be no way to map the Russian clusters to the Ukrainian ones, or vice

Table 2: Clustering correspondence to document topics

Method	**Incorrect assignments, %**
Mono-lingual	
Ukrainian	4.7
Russian	34.7
Cross-lingual	
Naive Binary	50.17
Naive Count	50.00
Edit distance translation Binary	50.50
Edit distance translation Count	50.50
Dictionary/Edit distance Binary	50.33
Dictionary/Edit distance Count	49.83
Matrix translation Binary	36.33
Matrix translation Count	36.17
Semantic fingerprints on word types	35.33
Semantic fingerprints on word tokens	**5.50**

versa.

So, the next step was to cluster all documents together, independent of their languages, using the techniques described in the Section 4.. The results are shown in the Table 2.

We used two simple baseline approaches. The first one is dubbed '**naive**': we cluster all the texts' bag-of-words representations as is, with no special preprocessing (only the PoS tags are unified across languages). Transformation from texts to bags-of-words resulted in 885-dimensional document vectors. This baseline approach exploits the intuition that in closely-related languages such as Russian and the Ukrainian there are many words which share both spelling and meaning. This is true, but this fact does not help *K-means* to correctly cluster the collection into topical classes: 50.17% of the documents are assigned an incorrect cluster, much more than in any of our mono-lingual experiments. Employing count vectors instead of binary ones lowers error rate only down to 50%. Using *tf-idf* weighting (Jones, 1972) did not significantly change the results neither for this nor for other baselines.

Looking into particular cluster assignments reveals that *K-means* clusters all the Ukrainian documents into one group, and then partitions Russian texts into two clusters roughly corresponding to history and everything else. This is quite expected: the Ukrainian alphabet contains several frequent characters missing in Russian ('ґ, є, і, ї'), while the Russian-specific characters ('ё, ъ, ы, э') are much rarer. Consequently, the Ukrainian documents contain a lot of Ukrainian words specific only to them, while Russian words (or their identically spelled Ukrainian counterparts) are used throughout the whole collection. Anyhow, '**naive**' approach can't adequately represent the topical structure of the test set.

The second baseline employs quazi-translation of Ukrainian words into Russian using the already described approach with Damerau-Levenshtein edit distance (Damerau, 1964). We replaced all words in the Ukrainian texts with the Russian words closest to them by edit distance. Only nouns, adjectives, verbs and abbreviations were replaced; replacements were selected only among the same part of speech as the original word, and in case of ties, target word with the highest frequency in Russian corpus was selected. Then the same bag-of-words representation (now 834-dimensional) was fed to the clustering algorithm. Though for many words '**edit distance translation**' works quite well, it did not help in clustering multilingual test set. Whether with binary or count vectors, *K-means* still grouped all the Ukrainian documents into one cluster, resulting in 50.5% of incorrect assignments. The possible reason is that there are still many incorrect 'Levenshtein translations' resulting in target entities which are correct Russian words, but never appear in Russian documents from our test set. This gives *K-means* the ground to separate the Ukrainian texts from all the other documents.

Then we experimented with translating Ukrainian words into Russian using the learned transformation matrix (**matrix translation**). For each Ukrainian word, we multiplied its vector in the Ukrainian model by the matrix and found the Russian word nearest to the resulting vector. Then the Ukrainian words were replaced with these 'translations' and the same bag-of-words document representations were constructed (resulting in 845-dimensional vectors). As a result, the *K-means* clustering moved substantially towards the intended topical grouping: only 36.33% of the documents were assigned incorrect clusters, and using count vectors made it 36.17%. In fact, the clustering algorithm correctly separated all history documents into one cluster independent of the language, while still mixing things up with law and economics (as we know, they are a bit more difficult to separate even in a monolingual setting). Thus, this document representation seems to be clearly superior to the baseline **naive** or **edit distance** approaches. It results in the documents grouping which is almost as efficient as mono-lingual clustering of Russian texts, but is still not on a par with Ukrainian mono-lingual clustering.

Note that these improvements cannot be explained by the sheer fact of employing a bilingual dictionary. We tried to use the same dictionary directly: that is, for the Ukrainian texts in the test set, replace all the words with their dictionary Russian translation. The remaining out-of-vocabulary words were 'translated' with the Dameral-Levenshtein distance approach. The results are reported in the Table 2 as **dictionary/edit distance** method. They are a bit better than the ones of the raw edit distance, but still far from the performance of the **matrix translation** method. It means that the algorithm itself is the cause of improvements.

Finally, the best results were received by employing the '**semantic fingerprint**' approach. Recall that this fingerprint is an average vector of all words in the document. Consequently, each document is represented with a 300-dimensional vector, supposedly reflecting its 'meaning'.

The average vector can be calculated either on vectors of **word types** or of **word tokens** (thus taking into account individual frequencies of words in the document). These two variants roughly correspond to **binary** and **count** variants of the previous methods, but we intentionally dub them in another way to emphasize that these representations

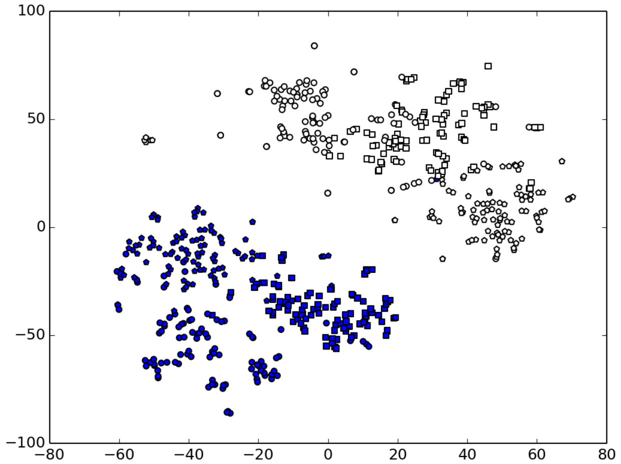


Figure 1: Naive baseline clustering

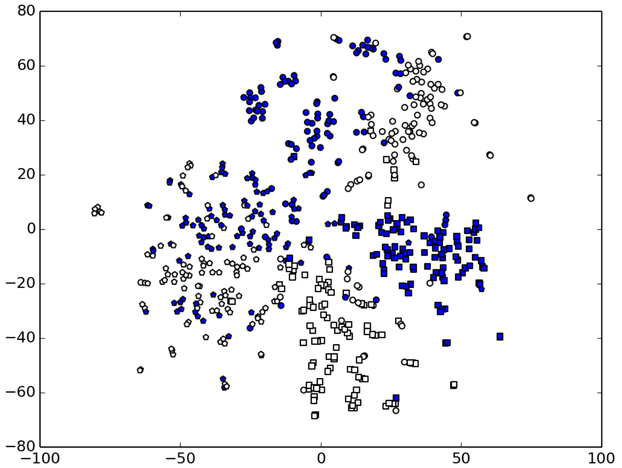


Figure 2: Matrix translation clustering

are radically different from bags-of-words. In this case we abstract away from particular words, and instead use some generalized 'semantic components', hopefully similar across languages.

We computed such fingerprints for all the documents in the test set, and for the Ukrainian documents we then multiplied the fingerprints by the transformation matrix, thus 'projecting' them into Russian semantic space. The resulting 300-dimensional representations are already numerical and can be directly fed into a clustering algorithm, without any bag-of-words preprocessing.

As a result, even rough semantic fingerprints calculated on word types (on **sets** of words in the documents) show clustering accuracy 1% better than the **matrix translation** approach. But as soon as semantic fingerprints are computed using word tokens (**lists** of words), the ratio of incorrect assignments drops drastically down to 5.5%. This result is very close to the quality of the mono-lingual Ukrainian clustering and is much better than that of the mono-lingual Russian clustering. It means that semantic fingerprints approach performs almost as good in the cross-lingual setting as traditional approaches in the mono-lingual one. Additionally, the fact that it outperformed the Russian monolingual clustering might mean that using dense vector representations for documents allowed to overcome the problems with separating economics and law texts in Russian, which seemed intractable for the bag-of-words approach.

To be more precise, into their respective clusters were grouped 196 of 200 economics documents (this corresponds to approximately 0.95 precision and 0.98 recall), 195 of 200 history documents (0.92 precision, 0.98 recall) and 176 of 200 law documents (0.97 precision, 0.88 recall), all independent of their languages. Total average F1 measure is about 0.95, which significantly outperforms the multilingual clustering performance reported in (Mathieu et al., 2004).

Figures 1, 2 and 3 illustrate clustering mechanics for the methods described above. We employed *t-SNE* dimensionality reduction technique (Van der Maaten and Hinton, 2008) to project high-dimensional representations[3] of the test set documents into 2-dimensional plots. Colors reflect document language (blue for Ukrainian and white for Russian), while marker types stand for document topic (circles for law, squares for history and pentagons for economics). Note that these projections inevitably lose a lot of information as compared to initial high-dimensional data, and should be considered as only approximate visualizations.

It is clearly visible that with **naive baseline** representations in the Figure 1 there are almost no links between different-language documents belonging to one topic. The dataset is clearly separated into Russian and Ukrainian clusters, and topics can be seen inside languages, but there is hardly a way to group documents into language-independent topical clusters. This is the reason for *K-means* failing to achieve our aim with the baseline approach. On the other hand, with **matrix translation** representations (Figure 2), language-independent topics already emerge, but still with much noise. Language boundaries are eroded, especially with economics documents.

Finally, with **semantic fingerprints** representations in the Figure 3 the structure of the test set is manifested in full. There are six well-defined clusters corresponding to topics and languages and a clear spatial structure, which allows *K-means* to easily group documents into 3 larger topical clusters without losing the ability to tell a Russian document from a Ukrainian one. Note how the Ukrainian topical clusters seem to share a common linear relation to the Russian ones, reminding about linear relations between different languages' vector spaces.

Thus, we were able to correctly cluster multilingual documents according to their topics without any proper 'translation' and without even considering word spelling. This means that, first, semantic fingerprints are precise enough to reveal topical differences between documents, and second, that this holds even after linear transformation of such fingerprints into another language semantic space.

6. Discussion

We tested 'transformed' semantic representations of the documents on the clustering task, but theoretically they can

[3] 300 dimensions for semantic fingerprints, 885 and 845 for naive baseline and matrix translation correspondingly.

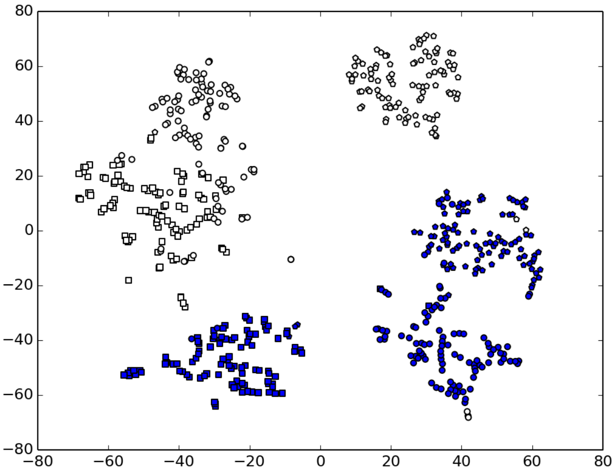


Figure 3: Semantic fingerprints clustering

be used for any problems which demand semantic-aware cross-lingual representations, including classification and visualization. Also, the number of involved languages is not limited in any way. The proposed method is relatively simple and straightforward to implement: one needs only comparable monolingual corpora to train CBOW models on them (using any of the available off-the-shelf toolkits) and a small bilingual dictionary for each language pair to train linear transformation matrices. After that, all the words and documents in the corpora can be transformed into a unified language-independent semantic representation.

It is interesting that in our experiments semantic fingerprints' performance was better than direct 'translation' of words using the same transformation matrix. It reveals an important advantage of such generalized representations: they do not depend on particular words. In the case of the bag-of-words approach, a small mistake in matrix translation can lead to replacement of a Ukrainian word with a Russian counterpart, which is semantically similar, but not exactly the one used to denote this concept in the Russian part of this text collection. As a result, this word becomes useless in representing documents cross-linguistically. On the other hand, with the semantic fingerprints approach, an 'approximate transformation' is enough, as it will still be close to the corresponding words from Russian texts in the vector space.

This also explains the accuracy boost this approach gets from considering word tokens instead of word types. Of course, one reason is that vectors for frequent (arguably more topical) words become more important in determining the final average value, but this is also the case for bag-of-words approaches. The difference is that with the latter (including 'matrix translation' method), frequencies of words from the same semantic field are interpreted as independent features. As a result, for example, n words of frequency z related to history topic will not be more important than other random n words of the same frequency. At the same time, with 'semantic fingerprints', topically connected words collectively increase or decrease expression of the corresponding semantic component or components: they are more important in determining the resulting fingerprint, than random noise words, even if they are frequent. This leads to better discrimination between documents of different topics.

It is also important that 'semantic fingerprinting' is significantly faster than 'matrix translation', as we eliminate the necessity to look for the most similar neighbors of the predicted vector. This operation can be computationally expensive, especially on models with large vocabulary.

Note that one can apply the described method not only to proper cross-lingual translations, but also to problems like 'projecting' texts in one style or genre into another. In fact the method is applicable to any situation, where there are two comparable sets of texts consisting of items, for which there theoretically exist pairwise links; one knows only a small part of these links, but would like to compare texts independent of the corpus they belong to. In these cases, semantic fingerprints method can be of use.

7. Conclusion and Future Work

Thus, we described an approach to build language-independent semantic representations of documents in multi-lingual comparable corpora. It was tested on a rather small task of clustering Ukrainian and Russian academic texts into 3 topics. However, the results seem very promising to us and we plan to continue working on the proposed method. The models trained on our corpora, the linear transformation matrix, the evaluation dataset we used and Python code to work with this data are available online[4].

The initial motivation behind this work was to develop a system for automatic plagiarism detection for two closely related languages. A crucial component of this system is a preprocessing part, which is able to cluster texts according to their topics. We believe that this component eventually will make it possible to compare 'semantic fingerprints' of the documents in order to determine possible plagiarized texts and to perform their further analysis.

One obvious disadvantage of the proposed method is the necessity to know in advance the desired number of clusters (topics in the text collection). We plan to experiment with approaches to determining the optimal number of clusters automatically. It poses serious problems in multi-lingual settings, as the algorithms will be biased to language-based clustering, not taking into account topical division. Thus, ways should be invented to cope with this bias especially when the number of topics is much higher than 3 (used in this research).

Another direction of future work is to compare our approach and bag-of-words representations after proper machine translation. The results are not obvious: on the one hand, MT directly casts texts into another language and that should be a difficult baseline to beat. On the other hand, as explained in Section 6., dense document representations like semantic fingerprints can possibly be more flexible in grasping document contents than words-based representations.

Finally, we plan to test the proposed method with other language pairs, especially typologically distant languages. Experiments in (Mikolov et al., 2013a) suggest that as long as

[4] https://cloud.mail.ru/public/Eune/tN7ssqtWj

the languages possess the meaningful notion of lexical co-occurrence, genetic or typologic distances between them should not matter. However, this is still to be tested and proved.

8. Acknowledgments

We are grateful to the voluntary *Dissernet* community for providing us with the data, support, and inspiration to start this work.

9. Bibliographical References

Coulmance, J., Marty, J.-M., Wenzek, G., and Benhalloum, A. (2016). Trans-gram, fast cross-lingual word-embeddings. *arXiv preprint arXiv:1601.02502*.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Commun. ACM*, 7(3):171–176, March.

Ganich, D. I. and Oleynik, I. S. (1990). *Russian-Ukrainian and Ukrainian Russian Glossary*. Veselka.

Hartigan, J. A. (1975). *Clustering algorithms*. Wiley.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Kotsyba, N., Mykulyak, A., and Shevchenko, I. V. (2009). Ugtag: morphological analyzer and tagger for the ukrainian language. In *Proceedings of the international conference Practical Applications in Language and Computers (PALC 2009)*.

Mathieu, B., Besançon, R., and Fluhr, C. (2004). Multilingual document clusters discovery. In *Coupling approaches, coupling media and coupling languages for information retrieval*, pages 116–125.

Mikolov, T., Le, Q., and Sutskever, I. (2013a). Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May.

Segalovich, I. (2003). A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In *MLMTA*, pages 273–280. Citeseer.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(2579-2605):85.

Wolf, L., Hanani, Y., Bar, K., and Dershowitz, N. (2014). Joint word2vec networks for bilingual semantic representations. *International Journal of Computational Linguistics and Applications*, 5(1):27–44.

A 2D CRF Model for Sentence Alignment

Yong Xu, François Yvon
LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay
{yong, yvon}@limsi.fr

Abstract

The identification of parallel segments in parallel or comparable corpora can be performed at various levels. Alignments at the sentence level are useful for many downstream tasks, and also simplify the identification of finer grain correspondences. Most state-of-the-art sentence aligners are unsupervised, and attempt to infer endogenous alignment clues based on the analysis of the sole bitext. The computation of alignments typically relies on multiple simplifying assumptions, so that efficient dynamic programming techniques can be used. Because of these assumptions, high-precision sentence alignment remains difficult for certain types of corpora, in particular for literary texts. In this paper, we propose to learn a supervised alignment model, which represents the alignment matrix as two-dimensional Conditional Random Fields (2D CRF), converting sentence alignment into a structured prediction problem. This formalism enables us to take advantage of a rich set of overlapping features. Furthermore, it also allows us to relax some assumptions in decoding.

Keywords: Sentence Alignment, Conditional Random Fields

1. Introduction

The extraction of parallel segments in parallel or comparable corpora can be performed at various levels of granularity (documents, paragraphs, sentence, phrases, chunks, words, etc). For parallel texts or *bitexts*, i.e. pairs of texts assumed to be mutual translations, sentence alignment is a well-defined task in the processing pipeline (Wu, 2010; Tiedemann, 2011). For comparable corpora, sentence alignment techniques are used to mine parallel segments (Munteanu and Marcu, 2005; Uszkoreit et al., 2010). Sentence alignment is used in many applications, such as Statistical Machine Translation (SMT) (Brown et al., 1991), Computer-Assisted Tools, Translator Training (Simard et al., 1993a) and Language Learning (Nerbonne, 2000; Kraif and Tutin, 2011). In SMT, sentence alignment mostly aims at extracting parallel sentence pairs from large-scale corpora (e.g. bilingual parliament proceedings, web-crawled multilingual materials) to fuel downstream statistical processing. For such use, the alignment problem is considered to be solved: on the one hand, it is possible to discard unreliable alignments or difficult pairs (although, as pointed out by Uszkoreit et al. (2010), this might lead to a waste of training material); on the other hand, Goutte et al. (2012) showed that the translation quality of SMT (as measured by BLEU and METEOR) is robust to noise levels of $\approx 30\%$ in sentence alignments.

For other applications, the situation is quite different: First, a requirement may be to align the full bitext, for instance in translation checking (Macklovitch, 1994) or bilingual reading (Pillias and Cubaud, 2015; Yvon et al., 2016). Second, certain types of corpora exhibit important translational irregularities, making high precision alignment difficult. In particular, Yu et al. (2012; Lamraoui and Langlais (2013) showed the link-level F-score of state-of-the-art sentence aligners on bilingual fictions remains unsatisfactory. It was for instance found that the best link-level F-score obtained for “De la Terre à La Lune” (J. Verne), a subpart of the BAF corpus (Simard, 1998), was only around 78%.

In this paper, we consider the full sentence alignment problem for difficult bitexts, e.g. literary works and study how supervised learning techniques can help improve this state of affair. More precisely, inspired by the approach of (Mújdricza-Maydt et al., 2013), we propose to represent the alignment matrix by a two-dimensional CRF model, supervised by both reference alignments and external parallel corpora. We use a binary variable to represent the existence of alignment relation between each source and target sentence pair. Once all variables are predicted, we can recover conventional alignment links from the posterior matrix. This representation is very general and dispenses with problematic assumptions, at the cost of a more complex inference procedure.

The rest of this paper is structured as follows. In Section 2., we review some state-of-the-art methods, analyze their limitations and motivate our model. We detail the training and inference in Section 3. Experiments are reported in Section 4. Finally, we conclude and give perspectives for future work in Section 5.

2. Motivations

The development of bitext sentence alignment techniques dates back to the early 90s (Brown et al., 1991; Gale and Church, 1991; Simard et al., 1993b; Chen, 1993). Thanks to a sustained research effort, many high-quality aligners are nowadays publicly available, see e.g. (Moore, 2002; Varga et al., 2005; Braune and Fraser, 2010; Lamraoui and Langlais, 2013). A recent evaluation of these tools is in (Xu et al., 2015).

Most state-of-the-art aligners share a two-step approach.[1] A first, relatively coarse decoding pass extracts a set of parallel sentence pairs that the system deems reliable (for instance using length-based information). These pairs serve as either anchor points to reduce the search space of subsequent steps, or as seeds to obtain better parallelism estimation tools (for instance a classifier or a bilingual lexicon), or both. A second decoding pass, using the information gathered during the first step, realigns the bitext. Most of these alignments tools are unsupervised, so that the system has

[1] (Melamed, 1999) is a notable exception.

to collect information from the sole bitext(s) that need to be aligned. In decoding, aligners often make the following assumptions: (a) alignment links lie around the bitext diagonal; (b) there exist limited number of link types. These two assumptions, together with the convention that alignment links are monotone and associate continuous spans,[2] warrant the use of dynamic programming (DP) techniques to perform the search. The resulting alignment tools are often light-weight and efficient, a major requirement if one wishes to process very large bitexts.

Despite their efficiency and good empirical performance on many corpora, existing sentence alignment tools suffer from a number of problems:

- probabilistic alignment models typically assume a fixed prior distribution over link types, as well as specific choices for length distributions (e.g. Gaussian or Poisson). However, Wu (1994) demonstrated that these assumptions could be inaccurate, especially for language pairs that are not closely related;

- as shown in (Yu et al., 2012), DP-based methods often give poor results for *null links*, i.e. links for which one side is empty. Among the five methods compared in this study, only (Melamed, 1999) predicted a similar number of null links as the reference, while others tended to miss a significant portion of them. A possible reason for this problem is the lack of a coherent scoring mechanism which would allow to fairly compare null and non-null links; this especially applies to methods using lexical clues;

- probabilistic alignment models rely on *local* features, and ignore contextual evidences. It might be beneficial to explore structural dependencies in the training;

- the limitation on link types is also overly restrictive. Six main link types are used in most studies: 0:1, 1:0, 1 : 1, 2 : 1, 1 : 2, 2 : 2, and it is a fact that these types rassemble a large majority of links for most text genres. Xu et al. (2015) however report that, in a reference corpus composed of partial sentence alignments for seven literary bitexts, the other types account for approximately 5% of the total number of links, a non-negligible portion for full-text alignment tasks. Besides, such intrinsic model errors can propagate during the DP process.

Inspired by the model of Mújdricza-Maydt et al. (2013), we propose a two-dimensional CRF model for sentence alignment. We use a binary variable to model the existence of the parallelism relation between one source-to-target sentence pair, and include contextual information in our predictions. Decoding consists of classifying each variable as negative or positive. Furthermore, the model structure is richer than that of Mújdricza-Maydt et al. (2013) and includes an explicit representation of null links.

3. The 2D CRF Model

3.1. The model

Given a sequence of source language sentences $E_1^I = E_1, ..., E_I$ and a sequence of target sentences $F_1^J = F_1, ..., F_J$,[3] we propose a 2D CRF model to predict the presence of link between any pair of sentences $[E_i; F_j]$, where $1 \leq i \leq I, 1 \leq j \leq J$. Note that similar models have also been developed for sub-sentential alignments (Niehues and Vogel, 2008; Cromières and Kurohashi, 2009; Burkett and Klein, 2012). Each pair $[E_i; F_j]$ gives rise to a binary variable $\mathbf{y}_{i,j}$, whose value is 1 (*positive*) if E_i is aligned to F_j, and 0 (*negative*) otherwise. For the sequence pair E_1^I and F_1^J, there are $I \times J$ such variables, collectively denoted as $\mathbf{y}$. Dependencies between links are modeled as follows. For each pair $[E_i; F_j]$, we assume that the associated variable $\mathbf{y}_{i,j}$ depends on $\mathbf{y}_{i-1,j}$, $\mathbf{y}_{i+1,j}$, $\mathbf{y}_{i,j-1}$, $\mathbf{y}_{i,j+1}$, $\mathbf{y}_{i-1,j-1}$ and $\mathbf{y}_{i+1,j+1}$. In other words, it depends on the presence of links $[E_{i-1}; F_j]$, $[E_{i+1}; F_j]$, $[E_i; F_{j-1}]$, $[E_i; F_{j+1}]$, $[E_{i-1}; F_{j-1}]$ and $[E_{i+1}; F_{j+1}]$. Figure 1 displays a graphical representation of the model.

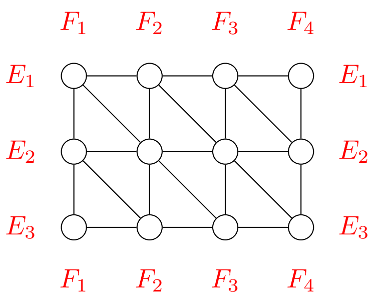


Figure 1: The 2D CRF model, for a bitext of 3 source $E_1 - E_3$ and 4 target sentences $F_1 - F_4$.

The topology of our model differs from the proposal of Mújdricza-Maydt et al. (2013), where each diagonal of the alignment matrix was modeled as a linear chain CRF. This topology captured the important diagonal direction dependency, but did not encode the horizontal or vertical dependencies. Another important difference lies on the generation of final outputs. Mújdricza-Maydt et al. (2013) variable labels to encode the corresponding link type (e.g. 1:1, 2:1). Note that this encoding makes it impossible to include all link types, and has also a bearing on the computational cost, since the inference complexity of a linear chain CRF is quadratic in the number of labels. As a result, these authors only considered 6 link types (1:1, 1:2, 2:1, 1:3, 3:1, F).[4] In our model, all prediction variables are binary. We generate final links using the transitive closure operation according to sentence alignment conventions, which can theoretically lead to any possible link type. For instance, an all zero-valued j^{th} column indicates an unaligned target sentence F_j; if $\mathbf{y}_{p,q}$ is the only positive value in the p^{th} row and q^{th}

[2] If a group of sentences on one side has no correspondence on the other side, they form a null link.

[3] Note we are referring languages as *source* and *target* only for convenience. Thus, *source* does not necessary indicate the original language of the bitext, nor does *target* indicate the translation language.

[4] F stands for all other link types or unaligned sentences.

column, then there is a 1:1 link $[E_p; F_q]$, etc. In fact, the model can express finer correspondences than conventional alignment link representations. For example, if both E_{u-1} and E_u are aligned to F_{v-1}, E_u is further aligned to F_v, our formalism can represent exactly the relation, while the alignment link representation would contain a coarser 2:2 link $[E_{u-1}, E_u; F_{v-1}, F_v]$.

We use two kinds of clique potentials in our model: *node potentials* and *edge potentials*. We impose that all single node cliques use the same clique template, i.e. they share the same set of feature functions and corresponding weights. For edge potentials, we use distinct clique templates for vertical, horizontal and diagonal edges. One main limitation of this model is that it does not include long distance dependencies, which makes it difficult to encode certain types of constraints (e.g. that alignment links should not cross). The model for a pair of sentence sequences $[E; F]$ (as a shorthand for $[E_1^I; F_1^J]$) can be written as:

$$p(\mathbf{y}|E,F) = \frac{1}{Z(E,F)} \prod_{\nu} \Phi_n(\mathbf{y}_\nu)\Phi_v(\mathbf{y}_\nu)\Phi_h(\mathbf{y}_\nu)\Phi_d(\mathbf{y}_\nu)$$

where $\nu \in \{(i,j) : 1 \leq i \leq I, 1 \leq j \leq J\}$, $\Phi_n(\mathbf{y}_\nu)$ stands for the single node potential at ν, $\Phi_v(\mathbf{y}_\nu)$ represents the potential on the vertical edge connecting ν and the node just below it:

$$\forall j, \Phi_v(\mathbf{y}_{i,j}) = \begin{cases} 1 & \text{if} \quad i = I \\ \Phi_v(\mathbf{y}_{i,j}, \mathbf{y}_{i+1,j}) & \text{if} \quad 1 \leq i < I \end{cases}$$

$\Phi_h(\mathbf{y}_\nu)$ (the horizontal potential) and $\Phi_d(\mathbf{y}_\nu)$ (the diagonal potential) are defined similarly. $Z(E,F) = \sum_{\mathbf{y}'} \prod_\nu \Phi_n(\mathbf{y}'_\nu)\Phi_v(\mathbf{y}'_\nu)\Phi_h(\mathbf{y}'_\nu)\Phi_d(\mathbf{y}'_\nu)$ is the normalization factor (*the partition function*) of the CRF. All potentials take the generic form of a log-linear combination of feature functions:

$$\Phi_\nu(\mathbf{y}_\nu) = \exp\{\boldsymbol{\theta}^\top \mathbf{F}_\nu(\mathbf{y}_\nu)\},$$

where $\mathbf{F}_\nu$ and $\boldsymbol{\theta}$ are the feature and weight vectors. We also use ℓ^2 regularization with scaling parameter $\alpha > 0$.[5]

3.2. Learning the 2D CRF model

The conventional learning criteria for CRF is the Maximum Likelihood Estimation (MLE). For a set of fully observed training instances $\mathcal{A} = \{(E^{(s)}, F^{(s)}, \mathbf{y}^{(s)})\}$, MLE consists of maximizing the log-likelihood of the training set with respect to model parameters $\boldsymbol{\Theta} = \{\boldsymbol{\theta}, \alpha\}$. The log-likelihood is concave with respect to the weight vector, which warrants the use of convex optimization techniques to obtain parameter estimates. In order to do this, we need the gradient of the likelihood function with respect to the weight vector.

Computing the gradients requires two kinds of marginal probabilities: single node marginals $p(\mathbf{y}_{i,j}|E^{(s)}, F^{(s)})$ and edge marginals $p(\mathbf{y}_{i,j}, \mathbf{y}_{i+1,j}|E^{(s)}, F^{(s)})$, $p(\mathbf{y}_{i,j}, \mathbf{y}_{i,j+1}|E^{(s)}, F^{(s)})$, and $p(\mathbf{y}_{i,j}, \mathbf{y}_{i+1,j+1}|E^{(s)}, F^{(s)})$. We need to perform inference to compute these marginals. Since the topology of our model contains loops, we use the *Loopy Belief Propagation* (LBP) inference algorithm. Even though LBP is an *approximate* inference algorithm with no convergence guarantee, Murphy et al. (1999) observe that it often gives reasonable estimates (assuming it converges).

For a tree-structured undirected graphical model, the message from a node $\mathbf{y}_\mu$ to a neighboring node $\mathbf{y}_\nu$ takes the following form (Wainwright and Jordan, 2008):

$$m_{\mu\nu}(\mathbf{y}_\nu) \propto \sum_{\mathbf{y}_\mu} \Phi(\mathbf{y}_\mu)\Phi(\mathbf{y}_\nu, \mathbf{y}_\mu) \prod_{\gamma \in N(\mu)\setminus\nu} m_{\gamma\mu}(\mathbf{y}_\mu)$$

where $N(\mu)$ denotes the set of neighbors of μ. LBP is performing such message passing procedure on a cyclic graph. Once message passing has converged, the single node and edge marginals (a.k.a. "beliefs") are expressed as:

$$b_\nu(\mathbf{y}_\nu) \propto \Phi(\mathbf{y}_\nu) \prod_{\gamma \in N(\nu)} m_{\gamma\nu}(\mathbf{y}_\nu)$$

$$b_{\mu\nu}(\mathbf{y}_\mu, \mathbf{y}_\nu) \propto \Phi(\mathbf{y}_\mu)\Phi(\mathbf{y}_\nu)\Phi(\mathbf{y}_\nu, \mathbf{y}_\mu) \prod_{\delta \in N(\mu)\setminus\nu} m_{\delta\mu}(\mathbf{y}_\mu) \prod_{\gamma \in N(\nu)\setminus\mu} m_{\gamma\nu}(\mathbf{y}_\nu)$$

In practice, it is possible that LBP does not converge for certain training instances. In this case, we simply stop it after 100 iterations. Convex optimization routines also require to compute the log-partition function $\log Z(E,F)$, as a part of the likelihood function. LBP approximates this quantity with the Bethe Free Energy (Yedidia et al., 2001).

In learning, we first train the CRF without any edge potential (thus making the model similar to the simpler MaxEnt model), and use it to initialize the parameter vector of node potentials. We then randomly initialize other parameters,[6] and use the L-BFGS algorithm (Liu and Nocedal, 1989) implemented in the SciPy package to perform parameter learning, this time with all potentials.

3.3. Search in the 2D CRF model

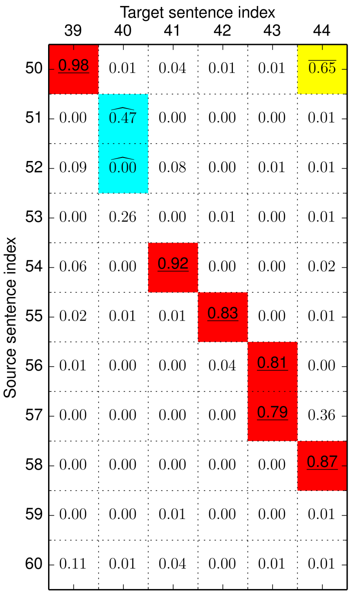
For the 2D CRF model, we perform the search in multiple steps. First, we run the BMA algorithm (Moore, 2002) to extract high-confidence 1:1 links. This algorithm first extracts reliable 1:1 sentence pairs from a bitext, using only length information, then trains a small IBM Model 1 based on these links, finally realigns the bitext using both length and lexical information. It returns a set of $1:1$ sentence pairs. As reported in (Yu et al., 2012), BMA tends to obtain a very good precision, at the expense of a less satisfactory recall. Furthermore, BMA computes posterior probabilities for every possible link, which are then used as confidence scores. We filter the result links with a very high posterior probability threshold (≥ 0.99999) (this threshold is much higher than BMA's default choice). These links segment the entire search space into sub-blocks. For each sub-block, we construct a 2D CRF model, and perform decoding. As exact Maximum A Posteriori decoding is intractable, instead, we run max-product LBP independently, and pick the *local best* label for each node. The label assigned to a variable $\mathbf{y}_{i,j}$ is

$$\arg\max_{l \in \{0,1\}} b_{i,j}(l)$$

[5] In the experiments, α is tuned on a development set, and takes the value 0.1.

[6] See (Sutton, 2008, 88–89) for a discussion on parameter initialization of general CRFs trained using LBP.

This procedure returns a set of *sentence-level* links. Since the sizes of the sub-blocks are often small (generally smaller than 10 × 10), decoding is very fast in practice.
Figure 2 displays an *alignment prediction matrix*.[7] It contains four types of cells, corresponding to four types of predictions: true positive (red, with underlined score), true negative (white, with normal score), false positive (yellow, with overlined score), false negative (cyan, with hatted score). The score in each cell is the marginal probability of the pair being positive, as computed by the CRF. A red or yellow cell indicates a sentence-level link predicted by the model.



	39	40	41	42	43	44
50	$\underline{0.98}$	0.01	0.04	0.01	0.01	$\overline{0.65}$
51	0.00	$\widehat{0.47}$	0.00	0.00	0.00	0.01
52	0.09	$\widehat{0.00}$	0.08	0.00	0.01	0.01
53	0.00	0.26	0.00	0.01	0.00	0.01
54	0.06	0.00	$\underline{0.92}$	0.00	0.00	0.02
55	0.02	0.01	0.01	$\underline{0.83}$	0.00	0.01
56	0.01	0.00	0.00	0.04	$\underline{0.81}$	0.00
57	0.00	0.00	0.00	0.00	$\underline{0.79}$	0.36
58	0.00	0.00	0.00	0.00	0.00	$\underline{0.87}$
59	0.00	0.00	0.01	0.00	0.00	0.01
60	0.11	0.01	0.04	0.00	0.01	0.01

Figure 2: An alignment prediction matrix.

Two types of errors exist in the alignment prediction matrix: false negatives (cyan cells with hatted scores) and false positives (yellow cells with overlined scores). We cannot easily deal with false negatives. False positives introduce noises, for example, the pair $(50, 44)$ in Figure 2 (the upper right corner). The two positive pairs $(50, 39)$ and $(50, 44)$ lead to two separate links involving the same source sentence, which violates the general convention of sentence alignment. In fact, the pair $(50, 44)$ is clearly wrong: it links the first source sentence with the last target one, thus overlapping with all other positive sentence-level links.[8] In our experiments, in all sub-blocks, true positive sentence-level links always lie around the main diagonals. We have used the following heuristics to smooth the alignment prediction matrix:

1. perform a linear regression on all predicted positive sentence-level links, then take a band of fixed width around the regression line, and drop positive links that lie outside of this band.[9]

2. if after this step, there are still separate links involving the same sentence, we take the positive sentence-level links in the surrounding window with width 5, and discard the ones which are inconsistent with the surrounding links;

3. if it is still undecidable, we perform again a linear regression of positive sentence-level links in the surrounding window, and discard the link that is farthest away from the regression line.

In practice, step 3 was hardly performed.
Finally, to turn sentence-level links into *alignment-level* links, we apply the following rules:

1. consecutive sentence-level links in the horizontal or vertical directions are combined into a large alignment-level link;

2. a sentence-level link without horizontal or vertical neighbors becomes a 1:1 type alignment-level link.

These rules follow from the interpretation of our model, where an $n{:}m$ type alignment-level link decomposes into $n * m$ sentence-level links.

4. Experiments

4.1. Features

In the 2D CRF model, feature functions take the form $f(E_1^I, F_1^J, i_1, j_1, i_2, j_2, \mathbf{y}_{i_1,j_1}, \mathbf{y}_{i_2,j_2})$, where E_1^I is the source sequence of I sentences, F_1^J the target sequence of J sentences, (i_1, j_1) and (i_2, j_2) neighboring source-target indices, $\mathbf{y}_{i_1,j_1}$ and $\mathbf{y}_{i_2,j_2}$ respectively corresponding labels (0 or 1). For each pair (E_i, F_j), we compute the following set of features:

1. The *length difference ratios*. We first compute

$$r_1 = \frac{|len(E_i) - len(F_j)|}{len(E_i)}, \; r_2 = \frac{|len(E_i) - len(F_j)|}{len(F_j)}$$

where the $len()$ function returns the number of characters in one string. Both r_1 and r_2 are rounded into the interval $[0, 1]$, then discretized into 10 indicator features. This family thus contains 20 features.

2. The *ratio of identical tokens*. Let the function $token()$ return the number of tokens in a string. We count the number of shared tokens in E_i and F_j, denote the count by s, compute two ratios $\frac{s}{token(E_i)}$ and $\frac{s}{token(F_j)}$, then discretize each into 10 features.

3. The *relative index difference*. We discretize the quantity $|\frac{i}{I} - \frac{j}{J}|$ into 10 features.

[7] Note these matrices are drawn just after the CRF decoding, before the post-processing described below.

[8] Note that this particular matrix was computed by an early version of the 2D CRF model. We show it here for illustration purpose. In later versions, the model is augmented with features capturing the relative position information, which effectively prevents this kind of errors.

[9] The band width is taken to be half of the number of sentences of the shorter one of the two sides.

Book	# Links	# Sent_EN	# Sent_FR
Alice's Adventures in Wonderland	746	836	941
Candide	1,230	1,524	1,346
Vingt Mille Lieues sous les Mers	778	820	781
Voyage au Centre de la Terre	714	821	754
Total	3,468	4,001	3,822

Table 1: The training corpus of the 2D CRF model.

Book	# Links	# Sent_EN	# Sent_FR
De la Terre à la Lune (BAF)	2,520	2,554	3,319
Du Côté de chez Swann	463	495	492
Emma	164	216	160
Jane Eyre	174	205	229
La Faute de l'Abbe Mouret	222	226	258
Les Confessions	213	236	326
Les Travailleurs de la Mer	359	389	405
The Last of the Mohicans	197	205	232
Total of `Manual en-fr`	1,792	1,972	2,102

Table 2: The test corpus, made of the literary part of BAF and the `manual en-fr` corpus.

4. The *lexical translation scores*. Let $token(E_i) = m$ and $token(F_j) = n$, we compute the IBM Model 1 scores:

$$\begin{aligned} T_1(E_i, F_j) &= \frac{1}{n}\sum_{s=1}^{n}\log(\frac{1}{m} * \sum_{k=1}^{m} p(F_{js}|E_{ik})) \\ T_2(E_i, F_j) &= \frac{1}{m}\sum_{k=1}^{m}\log(\frac{1}{n} * \sum_{s=1}^{n} p(E_{ik}|F_{js})) \end{aligned}$$

   where F_{js} is the s^{th} token of F_j. The lexical translation probabilities p are computed using an IBM 1 model trained on the EN-FR Europarl corpus (Koehn, 2005). After discretizing T_1 and T_2, we obtain 10 features for each alignment direction.

5. The *span coverage*. We split a string into several **spans** by segmenting on punctuations (except for the quotation marks). For each source span $span\_e$, we compute the translation score $T_2(span\_e, F_j)$. If the score is larger than a threshold,[10] we consider $span\_e$ as being **covered**. We then compute the ratio of covered source spans and the ratio of covered target spans, and discretize each into 10 features.

6. The *label transition*. These features capture the regularity of the transition of labels from one node (E_i, F_j) to one of its neighbors (e.g. (E_{i+1}, F_j)). For each of the three types of neighbors (vertical, horizontal, diagonal), we define four label transition features (because our prediction variables are binary). For example, for the vertical template, we define

$$\begin{aligned} g_{00}(i,j) &= \delta\{\mathbf{y}_{i,j} = 0 \wedge \mathbf{y}_{i+1,j} = 0\} \\ g_{01}(i,j) &= \delta\{\mathbf{y}_{i,j} = 0 \wedge \mathbf{y}_{i+1,j} = 1\} \\ g_{10}(i,j) &= \delta\{\mathbf{y}_{i,j} = 1 \wedge \mathbf{y}_{i+1,j} = 0\} \\ g_{11}(i,j) &= \delta\{\mathbf{y}_{i,j} = 1 \wedge \mathbf{y}_{i+1,j} = 1\} \end{aligned}$$

   where δ is the Kronecker delta function. We have similar features for horizontal and diagonal transitions. In total, this family contains 12 features.

7. The *augmented length difference ratio*. This family only applies to the vertical and horizontal edge potentials, under the condition that the two neighboring pairs are both positive. In the vertical (resp. horizontal) case, we combine the two consecutive source (resp. target) sentences E_i, E_{i+1} (resp. F_j, F_{j+1}) into one new sentence E' (resp. F'), then apply the computations carried out for feature family 1 for the pair (E', F_j) (resp. (E_i, F')).

8. The *augmented translation score*. This family only applies to vertical and horizontal edge potentials, under the condition that the two neighboring pairs are both positive. We construct E' (resp. F') as in the previous feature family. We then compute the augmented translation score $T_1(E', F_j) - T_1(E_i, F_j)$ (resp. $T_2(E_i, F') - T_2(E_i, F_j)$). The intuition is that a longer partial translation is better than a shorter one. Each score is discretized into 10 features.

Note feature families 6, 7 and 8 are computed only when possible. Feature families 5, 7 and 8 are new in our model. Others have been used in previous methods, for instance, (Munteanu and Marcu, 2005; Yu et al., 2012; Tillmann and Hewavitharana, 2013; Mújdricza-Maydt et al., 2013).

4.2. Learning corpus

The training of the 2D CRF model requires reference alignments. We have used the reference sentence alignments collected for an ongoing project.[11] The training corpus contains alignment links of four books: "Alice's Adventures in Wonderland" (L. Carroll), "Candide" (Voltaire), "Vingt

[10] In our experiments, the threshold is set to $\log(1e - 3)$

[11] See `http://transread.limsi.fr`, where most textual resources can be downloaded.

	Sentence level F-score						
	GMA	BMA	Hunalign	Garg	Yasa	MaxEnt	CRF
De la Terre à la Lune (BAF)	72.9	77.3	81.9	77.3	86.2	76.6	84.0
Du Côté de chez Swann	95.4	88.9	89.4	95.0	95.2	96.0	94.3
Emma	73.8	52.1	62.8	61.2	73.8	71.2	69.4
Jane Eyre	88.0	54.6	59.4	84.2	82.5	88.0	77.2
La Faute de l'Abbé Mouret	94.8	83.8	82.8	98.7	97.7	98.9	90.8
Les Confessions	82.8	49.9	48.5	80.5	82.8	86.1	76.6
Les Travailleurs de la Mer	87.8	79.6	78.8	91.5	90.4	91.9	89.1
The Last of the Mohicans	94.9	76.0	77.0	95.6	94.5	95.0	91.1
Average on `manual en-fr`	88.2	69.3	71.2	86.7	88.1	89.6	84.1

Table 3: Sentence level F-scores of the 2D CRF method on the test corpus, compared with state-of-the-art methods.

	BMA			MaxEnt			CRF		
	P	R	F	P	R	F	P	R	F
De la Terre à la Lune (BAF)	97.2	64.1	77.3	72.0	81.8	76.6	95.5	74.9	84.0
Du Côté de chez Swann	99.5	80.3	88.9	97.1	94.9	96.0	96.3	92.5	94.3
Emma	89.8	36.7	52.1	62.8	82.3	71.2	76.1	63.7	69.4
Jane Eyre	93.7	38.5	54.6	86.7	89.3	88.0	86.6	69.6	77.2
La Faute de l'Abbé Mouret	99.5	72.3	83.8	98.9	98.9	98.9	98.2	84.5	90.8
Les Confessions	98.4	33.4	49.9	89.3	83.2	86.1	92.6	65.4	76.6
Les Travailleurs de la Mer	97.7	67.2	79.6	90.8	93.0	91.9	97.1	82.2	89.1
The Last of the Mohicans	98.7	61.8	76.0	94.2	95.8	95.0	97.1	85.7	91.1
Average on `manual en-fr`	96.8	55.7	69.3	88.5	91.1	89.6	92.0	77.7	84.1

Table 4: The comparison of BMA, MaxEnt and the 2D CRF model, using sentence-level measures. P stands for Precision, R is Recall, and F is F-score.

Mille Lieues sous les Mers”, and “Voyage au Centre de la Terre” (both by J. Verne). Table 1 displays the statistics of the training corpus.

We have to convert the training corpus into a training set. A training instance is a fully observed sentence-level alignment matrix. In order to make train conditions as close as possible to test conditions, each fully aligned book was segmented into sub-blocks, again using high confidence 1:1 links computed by BMA as anchor points. Each sub-block, annotated with reference alignments, is then turned into one training instance. This strategy has the additional benefit to greatly reduce the total number of prediction variables, hence make the training less memory consuming. Besides, the training can enjoy better parallelization. There is potentially another advantage of using smaller training instances. Since our model contains one predictive variable for each pair of source-target sentences, there are roughly quadratically many negative examples, and linearly many positive ones. This data unbalance problem becomes more severe as the size of the prediction matrix grows larger. Using smaller training instances helps alleviate this problem.

Using this strategy, we obtain 450 fully observed alignment matrices. We use 360 for the training set, 90 as the development set. Among the 7,095 labeled sentence pairs, approximately 77% are negative.

For the test, we use the fully aligned novel “De la Terre à la Lune” in the BAF corpus and the `manual en-fr` corpus composed of 7 partial alignments of literary bitexts. Table 2 gives the statistics of the test corpus. Recall our first step is to use filtered results of BMA as anchor points to segment the search space. With the filtering threshold 0.99999, the anchor point precision is 0.89 on “De la Terre à la Lune”, and 0.96 on the `manual en-fr` corpus.

4.3. Results

We evaluate alignment results at two levels of granularity: the alignment level and the sentence level. At the alignment level, a link in the output alignment is considered correct if exact the same link is also in the reference alignment. At the sentence level, we decompose a $m{:}n$ type link in the reference alignment into $m \times n$ sentence pairs, all considered as correct. The same decomposition applies to computed links. We summarize precision and recall ratios into F-scores.

Since the 2D CRF model is intrinsically trained to optimize **sentence-level** metrics, we first look at its sentence-level performance, summarized in Table 3. For the sake of comparison, we also display the performance of six other state-of-the-art aligners: GMA (Melamed, 1999), BMA, Hunalign (Varga et al., 2005), Garg (as shorthand for Gargantua) (Braune and Fraser, 2010), Yasa (Lamraoui and Langlais, 2013), MaxEnt (Xu et al., 2015). The CRF model achieves great improvements over BMA and Hunalign. Its average score on the `manual en-fr` corpus is slightly inferior to other systems, but it obtains the second best F-measure on the large bi-text “De la Terre à la Lune”. We note that Yasa, perhaps the most lightweight tool, is very robust with respect to the sentence-level measure.

The first decoding step of both MaxEnt and CRF uses a subset of BMA's results as anchors to segment the bi-text space. Table 4 compares in more detail the performance of these three methods. (Yu et al., 2012; Lamraoui and Langlais, 2013) have reported that BMA usually delivers very high precision 1 : 1 links. We observe the 2D CRF model preserves a high sentence level precision, and greatly increases the recall. Thus, the 2D CRF model manages to extract true positive sentence pairs from the gaps defined by BMA's links with a very high accuracy. The behavior of MaxEnt varies on different corpus. On the `Manual en-fr` corpus, while it slightly decreases the precision, it obtains the best recall, leading to the best overall performance. However, on "De la Terre à la Lune", its precision is too low compared to BMA and CRF, thus its F-score is worse.

	Alignment level F-score		
	BMA	MaxEnt	CRF
De la Terre à la Lune (BAF)	73.6	66.5	73.3
Du Côté de chez Swann	91.5	93.3	90.9
Emma	57.4	51.0	55.4
Jane Eyre	61.1	78.9	63.2
La Faute de l'Abbé Mouret	88.4	98.0	82.8
Les Confessions	59.6	74.0	58.1
Les Travailleurs de la Mer	83.4	85.3	83.0
The Last of the Mohicans	82.7	90.1	84.3
Average on `manual en-fr`	74.9	81.5	74.0

Table 5: Alignment level F-scores of the 2D CRF model, compared with BMA and MaxEnt.

The alignment level F-scores of the CRF model are in Table 5.[12] The CRF achieves comparable alignment level F-scores to BMA on both sub-corpus. Although their average scores on `manual en-fr` are worse than MaxEnt, they outperform it considerably on those more difficult bitexts: "De la Terre à la Lune" and "Emma". In our opinion, this calls for further analyses for the *deployment* of alignment methods: for sentence alignment, it might be beneficial to investigate which types of methods tend to perform well for which types of bitexts, identify indicative characteristics (of methods and bitexts), and deduce operational guidelines.[13] Table 4 and Table 5 together show that, while the 2D CRF model obtains much higher sentence level F-scores than BMA (approximately 15 points on average on `manual en-fr`), their alignment level F-scores are actually comparable. In other words, the CRF does find more true positive sentence pairs, but not all of them contribute to form true links. Take for instance the 2 : 2 link $(14, 15; 24, 25)$ in Figure 3. To correctly recover this link, it is necessary to find at least three among the four cells. Even though the CRF finds one cell $(15; 25)$, this only yields a wrong 1 : 1 link, which, for the alignment level F-score metric, is no better than not finding any pair. While this imbalance between the alignment level and sentence level F-scores can seem surprising, it is by no means uncommon. In fact, this phenomenon was the reason that sentence-level F-score was proposed as an evaluation metric for sentence alignment in (Langlais et al., 1998). Nonetheless, this reinforces our belief that the deployment strategy of alignment methods, as well as evaluation metrics, needs further study.

4.4. Analysis

Error distribution by link type To better understand the behavior of the 2D CRF model, we perform an error analysis of its results on the `manual en-fr` corpus, with respect to link types. The corresponding statistics are in Table 7. We compare CRF with the MaxEnt approach, which gives the best average score on this corpus.

Link type	in Ref.	Error MaxEnt	Error CRF
0:1	20	18	15
1:0	21	18	15
1:1	1,366	105	64
1:2	179	36	**98**
1:3	32	9	**29**
2:1	96	32	**54**
2:2	24	19	20
others	27	15	26
Total	1,765	252	321

Table 7: Analyses of the errors of the MaxEnt and the CRF by link type, relative to the number of reference links (in Ref.), for the `manual en-fr` corpus. For example, 20 0 : 1 links are in the reference, and MaxEnt missed 18 of them. Only the link types occurring more than 5 times are reported. This filters out 27 links out of 1,792.

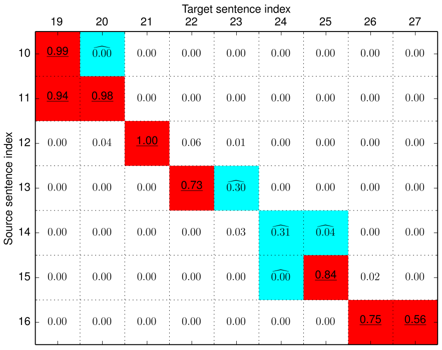


Figure 3: An alignment prediction matrix for a passage of "Les Confessions".

Compared to the MaxEnt method, CRF has a higher recall on null and 1:1 links. Its main weakness lies in the prediction of $1{:}n$ and $n{:}1$ links. After a closer study of the erroneous instances, we find a common pattern of error: when predicting a $m{:}n$ link with $m * n > 1$ (that is, a 1-to-many or many-to-many link), the CRF often correctly labels some sentence pairs as positive, while leaving others as negative. Figure 3 displays an alignment prediction matrix for a pas-

[12] We only show BMA, MaxEnt and CRF in this table, since (Xu et al., 2015) reported MaxEnt obtained the best average alignment F-score on the `manual en-fr` corpus.

[13] This is in line with the views of Deng et al. (2007) and Lamraoui and Langlais (2013), who suggested to model sentence alignment as part of the target application, so that it can benefit the optimization conducted toward the task.

	2D CRF			MaxEnt		
	#Null (in Ref.)	#Null (in Hyp.)	#Correct	#Null (in Ref.)	#Null (in Hyp.)	#Correct
De la Terre à la Lune (BAF)	714	1,311	672	714	150	91
Du Côté de chez Swann	9	27	8	9	5	3
Emma	41	85	28	41	2	2
Jane Eyre	10	77	7	10	0	0
La Faute de l'Abbé Mouret	2	52	2	2	1	1
Les Confessions	11	96	11	11	4	2
Les Travailleurs de la Mer	5	78	3	5	2	0
The Last of the Mohicans	12	37	3	12	2	2

Table 6: Performance of the 2D CRF model and the MaxEnt model on predicting null sentences. "#Null in Ref." is the number of unaligned sentences in the reference alignment; "#Null in Hyp." is the number of unaligned sentences in the hypothesis alignment computed by the model;"'#Correct" is the number of correctly predicted null sentences.

sage of Jean-Jacques Rousseau's "Les Confessions". The corresponding text (correctly aligned) is displayed in Table 8 in the appendix. The CRF fails to predict the 1:2 link $(13; 22, 23)$, only labelling $(13; 22)$ as positive; nor does it find the 2:2 link $(14, 15; 24, 25)$.

The failures of the 2D CRF model on 1-to-many and many-to-many links makes it necessary to study edge potentials. One of the reasons of using a CRF model is its ability to encode the dependencies between neighboring links, with which we expect to better predict non 1:1 links. An obvious direction to investigate is to add more edge features. Current edge features (families 6, 7 and 8) are quite general. It might be helpful to add features that encode finer level clues to edge potentials, e.g. word alignment information.

Besides of features of edge potentials, it might also be possible to consider other alignment matrix decoding algorithms. Compared to our approach, MaxEnt has the advantage of directly scoring alignment-level links, rather than doing it obliquely through sentence-level ones. This is also possible in the 2D CRF model, since LBP can readily compute marginals over edges, or even larger factors. We might use such marginals to improve our post-processing routines.

Null sentences Another motivation for the 2D CRF model is that it provides a mechanism where null and non-null links are handled coherently. We summarize its performance for null sentences in Table 6, again, comparing it with the MaxEnt method.

Although the 2D CRF model incorrectly labels many sentences as unaligned, it is indeed able to find the majority of true null sentences, except for "The Last of the Mohicans". This is where our model seems to be improving, especially when compared to MaxEnt.

5. Conclusion

In this paper, we reviewed state-of-the-art sentence alignment methods, identified several recurring problems, and have accordingly proposed a two-dimensional Conditional Random Fields model for the full text sentence alignment task. Our model is theoretically attractive, since it avoids several risky assumptions, computes posterior probabilities for all sentence alignment links, thereby explicitly representing null links, and warrants structured learning of parallelism scores.

In the light of our experimental results and analyses, we conclude that there is clear room of improvement for our 2D CRF model. Currently, while the model is effective at identifying true $1:1$ links with better recall than BMA's, its performance as measured by alignment level metric still needs to be improved. As perspectives, we would like to study the following improvements:

- enforce edge features: current edge features do not seem to be strong enough to balance our rich set of node features. Including features informed with simple word alignment information, such as fertilities and linked regions, seems an obvious way to go;
- add node features that encode the decisions of other systems, e.g. BMA;
- explore ways to simulate a DP process using marginals of edges or larger factors, which might help improve our alignment matrix decoding algorithm.

In the long term, we would like to study ways to characterize tasks and alignment methods, such that it is possible to choose adequate alignment algorithms for specific task requirements.

6. Acknowledgements

This work was partly supported by French National Research Agency under project TransRead (ANR-12-CORD-0015).

7. Bibliographical References

Braune, F. and Fraser, A. (2010). Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In *Coling 2010: Posters*, pages 81–89.

Brown, P. F., Lai, J. C., and Mercer, R. L. (1991). Aligning sentences in parallel corpora. In *Proceedings of ACL*, pages 169–176.

Burkett, D. and Klein, D. (2012). Fast inference in phrase extraction models with belief propagation. In *Proceedings of NAACL: HLT*, pages 29–38.

Chen, S. F. (1993). Aligning sentences in bilingual corpora using lexical information. In *Proceedings of ACL*, pages 9–16.

Cromières, F. and Kurohashi, S. (2009). An alignment algorithm using belief propagation and a structure-based distortion model. In *Proceedings of EACL*, pages 166–174.

Deng, Y., Kumar, S., and Byrne, W. (2007). Segmentation and alignment of parallel text for statistical machine translation. *Natural Language Engineering*, 13(03):235–260.

Gale, W. A. and Church, K. W. (1991). A program for aligning sentences in bilingual corpora. In *Proceedings of ACL*, pages 177–184.

Goutte, C., Carpuat, M., and Foster, G. (2012). The impact of sentence alignment errors on phrase-based machine translation performance. In *Proceedings of AMTA*.

Kraif, O. and Tutin, A. (2011). Using a bilingual annotated corpus as a writing aid: An application for academic writing for EFL users. In *Corpora, Language, Teaching, and Resources: From Theory to Practice. Selected papers from TaLC7*.

Lamraoui, F. and Langlais, P. (2013). Yet Another Fast, Robust and Open Source Sentence Aligner. Time to Reconsider Sentence Alignment? In *Proceedings of MT Summit*, pages 77–84.

Langlais, P., Simard, M., and Véronis, J. (1998). Methods and practical issues in evaluating alignment techniques. In *Proceedings of ACL-COLING*, pages 711–717.

Liu, D. C. and Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45:503–528.

Macklovitch, E. (1994). Using bi-textual alignment for translation validation: the TransCheck system. In *Proceedings of AMTA*, pages 157–168.

Melamed, I. D. (1999). Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25:107–130.

Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of AMTA*, Lecture Notes in Computer Science 2499, pages 135–144.

Mújdricza-Maydt, E., Köerkel-Qu, H., Riezler, S., and Padó, S. (2013). High-precision sentence alignment by bootstrapping from wood standard annotations. *The Prague Bulletin of Mathematical Linguistics*, (99):5–16.

Munteanu, D. S. and Marcu, D. (2005). Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Murphy, K. P., Weiss, Y., and Jordan, M. I. (1999). Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of UAI*, pages 467–475.

Nerbonne, J., (2000). *Parallel Texts in Computer-Assisted Language Learning*, chapter 15, pages 354–369. Text Speech and Language Technology Series.

Niehues, J. and Vogel, S. (2008). Discriminative word alignment via alignment matrix modeling. In *Proceedings of WMT*, pages 18–25.

Pillias, C. and Cubaud, P. (2015). Bilingual reading experiences: What they could be and how to design for them. In *Proceedings of INTERACT 2015*, pages 531–549.

Simard, M., Foster, G., and Perrault, F. (1993a). Transsearch: A bilingual concordance tool. Technical report, Centre for Information Technology Innovation.

Simard, M., Foster, G. F., and Isabelle, P. (1993b). Using cognates to align sentences in bilingual corpora. In *Proceedings of the 1993 Conference of the Centre for Advanced Studies on Collaborative Research*, pages 1071–1082.

Simard, M. (1998). The BAF: a corpus of English-French bitext. In *Proceedings of LREC*, pages 489–494.

Sutton, C. (2008). *Efficient Training Methods for Conditional Random Fields*. Ph.D. thesis, University of Massachusetts.

Tiedemann, J. (2011). *Bitext Alignment*. Number 14 in Synthesis Lectures on Human Language Technologies, Graeme Hirst (ed).

Tillmann, C. and Hewavitharana, S. (2013). A unified alignment algorithm for bilingual data. *Natural Language Engineering*, 19:33–60.

Uszkoreit, J., Ponte, J., Popat, A., and Dubiner, M. (2010). Large scale parallel document mining for machine translation. In *Proceedings of COLING*, pages 1101–1109.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., and Nagy, V. (2005). Parallel corpora for medium density languages. In *Proceedings of RANLP*, pages 590–596.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, January.

Wu, D. (1994). Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *Proceedings of ACL*, pages 80–87.

Wu, D. (2010). Alignment. In *CRC Handbook of Natural Language Processing*, number 16, pages 367–408.

Xu, Y., Max, A., and Yvon, F. (2015). Sentence alignment for literary texts. *Linguistic Issues in Language Technology*, 12(6).

Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2001). Generalized belief propagation. In *Proceedings of NIPS*, pages 689–695.

Yu, Q., Max, A., and Yvon, F. (2012). Revisiting sentence alignment algorithms for alignment visualization and evaluation. In *Proceedings of BUCC*, Istanbul, Turkey.

Yvon, F., Xu, Y., Pillias, C., Cubaud, P., and Apidianaki, M. (2016). Transread: Designing a bilingual reading experience with machine translation technologies. In *Proceedings of NAACL'16 (demo session)*.

8. Language Resource References

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *MT summit*, 5:79–86.

Appendix

Table 8 contains the text of a passage of Jean-Jacques Rousseau's "Les Confessions", corresponding to the alignment prediction matrix in Figure 3.

en$_{10}$ en$_{11}$	My mother's circumstances were more affluent; she was daughter of a Mons. Bernard, minister, and possessed a considerable share of modesty and beauty; indeed, my father found some difficulty in obtaining her hand.	Ma mère, fille du ministre Bernard, était plus riche: elle avait de la sagesse et de la beauté. Ce n'était pas sans peine que mon père l'avait obtenue.	fr$_{19}$ fr$_{20}$
en$_{12}$	The affection they entertained for each other was almost as early as their existence; at eight or nine years old they walked together every evening on the banks of the Treille, and before they were ten, could not support the idea of separation.	Leurs amours avaient commencé presque avec leur vie; dès l'âge de huit à neuf ans ils se promenaient ensemble tous les soirs sur la Treille; à dix ans ils ne pouvaient plus se quitter.	fr$_{21}$
en$_{13}$	A natural sympathy of soul confined those sentiments of predilection which habit at first produced; born with minds susceptible of the most exquisite sensibility and tenderness, it was only necessary to encounter similar dispositions; that moment fortunately presented itself, and each surrendered a willing heart.	La sympathie, l'accord des âmes, affermit en eux le sentiment qu'avait produit l'habitude. Tous deux, nés tendres et sensibles, n'attendaient que le moment de trouver dans un autre la même disposition, ou plutôt ce moment les attendait eux-mêmes, et chacun d'eux jeta son coeur dans le premier qui s'ouvrit pour le recevoir.	fr$_{22}$ fr$_{23}$
en$_{14}$ en$_{15}$	The obstacles that opposed served only to give a decree of vivacity to their affection, and the young lover, not being able to obtain his mistress, was overwhelmed with sorrow and despair. She advised him to travel – to forget her.	Le sort, qui semblait contrarier leur passion, ne fit que l'animer . Le jeune amant ne pouvant obtenir sa maîtresse se consumait de douleur: elle lui conseilla de voyager pour l'oublier .	fr$_{24}$ fr$_{25}$
en$_{16}$	He consented – he travelled, but returned more passionate than ever, and had the happiness to find her equally constant, equally tender.	Il voyagea sans fruit, et revint plus amoureux que jamais. Il retrouva celle qu'il aimait tendre et fidèle.	fr$_{26}$ fr$_{27}$

Table 8: The correct alignment of a passage of Jean-Jacques Rousseau's "Les Confessions", corresponding to the alignment prediction matrix in Figure 3.

Parallel Document Identification using Zipf’s Law

Mehdi Mohammadi
Department of Computer Science
Western Michigan University, MI, USA
mehdi.mohammadi@wmich.edu

Abstract

Parallel texts are an essential resource in many NLP tasks. One main issue to take advantage of these resources is to distinguish parallel or comparable documents that may have parallel fragments of texts from those that have no corresponding text. In this paper we propose a simple and efficient method to identify parallel documents based on Zipfian frequency distribution of available parallel corpora. In our method, we introduce a score called $CumulativeFrequencyLog$ by which we can measure the similarity of two documents that fit into a simple linear regression model. The regression model is generated based on the word ranks and frequencies of an available parallel corpus. The evaluation of the proposed approach over three language pairs achieve accuracy up to 0.86.

Keywords: Parallel corpora, Comparable Corpora, Parallel document identification, Zipf’s Law, Wikipedia.

1. Introduction

Statistical NLP approaches, such as Statistical Machine Translation (SMT), are highly attractive and yield satisfactory results. However, a prerequisite for such methods is a parallel corpus containing a large amount of correct translation pairs i.e. sentences in the source language aligned with their translations in the target language. Constructing parallel corpora for scarce resource languages is an expensive job, since it requires translators who are fluent in both source and target languages. It also takes a lot of time to collect such examples. Therefore, researchers have paid attention to some other online sources like bilingual web sites to create parallel corpora.

Zipf’s law is a statistical formulation devised empirically by G. K. Zipf that says in a corpus of natural language tokens, the frequencies of words associate inversely with their rank. This implies that rank-frequency distribution of words falls into an inverse relation. Two parallel corpora have this characteristic in common, so the frequency distribution of the words in one corpus would estimate the frequency of the words in the other side. In other words, the rank and frequency distribution of the terms in both documents are very close to each other.

In this paper we propose a method to identify parallel documents using a heuristic method based on Zipf’s law. The essence of the filter is based on Zipfian frequency distribution of two parallel corpora combined with a linear regression model. The linear regression model is obtained from frequency analysis of tokens in the parallel corpora. Zipf’s filter determines if two documents should be considered parallel or not using the error of prediction of linear regression function.

The motivation behind this work is to prepare fast and easy-to-build parallel corpora for limited-resource languages like Maori (the native language of New Zealand) to be used in NLP-related tasks. Beyond Statistical Machine Translation, such parallel corpora can be used in dialect identification (Malmasi et al., 2015) or lexicon construction. The proposed approach can also be extended to other NLP applications that deal with parallel corpus such as cross-language plagiarism detection in which a suspicious document is highly correlated to the original document in terms of words frequency distribution.

A primary application of this method is to find parallel documents among a set of comparable documents. Another interesting use case would be identifying comparable articles in Wikipedia and extracting parallel fragments of text from those comparable articles. Wikipedia is a source of multilingual texts that can be used to extract bilingual phrases or sentences automatically. Extracted parallel texts have been used as a complementary resource to Statistical Machine Translation systems in order to improve the performance of translation (Pal et al., 2014). Each article in Wikipedia may have a link to other languages. So, Wikipedia articles are aligned at document level. But they are not necessarily translations of each other. Although the articles with the same title in different languages are not exact translations of each other, it is possible to extract chunks of texts that have corresponding translations.

The rest of this paper is organized as follows. Section 2. presents an overview of the current approaches in this field. Section 3. presents details to undertake Zipf’s filter for parallel documents identification. In section 4. we show our experimental results and evaluations. Finally we conclude the paper in section 5.

2. Related Work

There are many attempts to align parallel texts at document level. Among the existing approaches, heuristic methods have been shown to be attractive and efficient for identifying comparable and parallel documents. The main advantage of these methods is that they are usually easy to implement as well as easy to understand.

The work in (Paramita et al., 2013) reports implementing two simple filters to detect comparable documents in Wikipedia articles. These filters are document’s minimum size and length’s difference. Using these filters they rule out over 80% of the initial document pairs.

Zafarian et al. (2015) use different characteristics of German-English documents in four modules to identify their similarity. These modules perform reducing the size of target space, Name Entity recognition, building topic

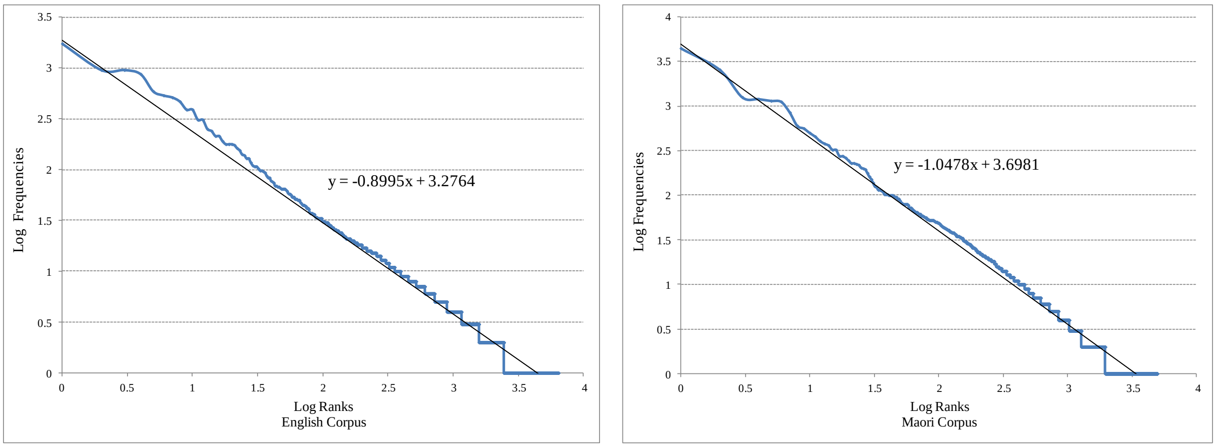


Figure 1: Zipf's curve for words in (a) English side, and (b) Maori side of a parallel corpus.

Parameters	**English**	**Maori**
Number of sentences	1695	1695
Number of words	30130	39488
Number of unique words	6380	4939

Table 1: The statistics of a small-size parallel corpora to analyze Zipf's law characteristics.

models and SMT. Their approach uses content of documents without links, tags or meta-data. Their results show that their approach can achieve recall of 45% for the first match.

In a system called LINA (Morin et al., 2015), authors use counts of hapax words to identify comparable documents. In their approach, only words that have appeared once in the document are considered for comparability measurement. Two documents that share the largest number of these hapax words are identified as parallel. Their results indicate that the system finds comparable documents with a precision of about 60%.

3. Zipfian-based Filter

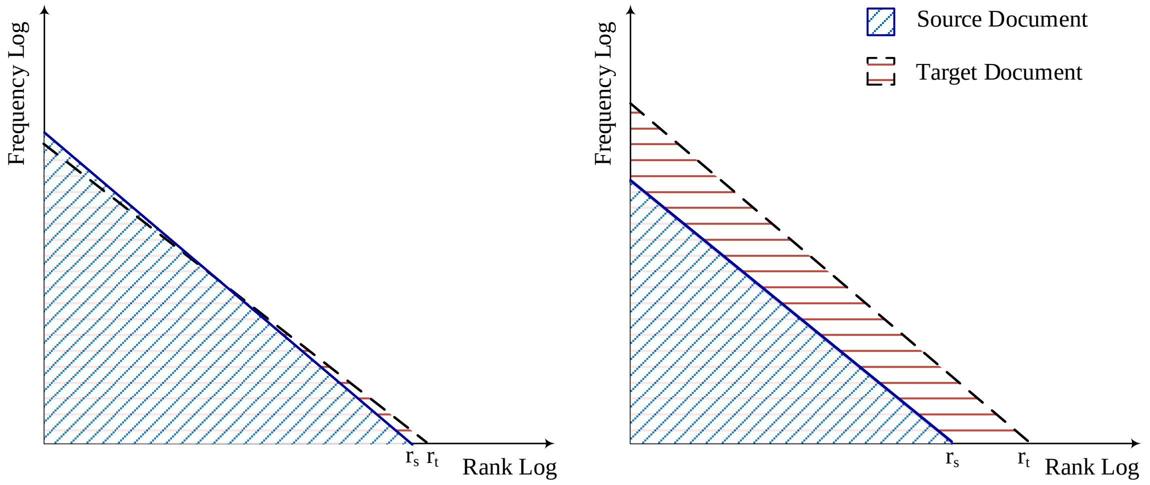
Based on Zipf's law, the frequency of words in a large corpus is inversely proportional to their rank (Deane, 2005). The empirical law for single word frequency distribution says that if the words in a corpus are ranked by their frequency, for a given word with rank r, the function $f(r)$ gives the frequency of the word such that

$$f(r) = \frac{C}{r^{\alpha}} \tag{1}$$

where C is a normalizing constant for the corpus and α is a free parameter for specifying the degree of skew. For single word frequency distribution, α is close to 1. The study by Ha et al. (2002) shows that beyond one token, a list of n-gram tokens also follow the law very well. Putting the logarithm of frequencies versus the logarithm of the ranks in a graph, a straight-like curve is obtained with slope of -1. For large corpora with about one million tokens, it has been observed that the highest ranked words may have frequencies that deviated slightly from the straight line. However, it is asserted that the law is valid for small corpora (Ha et al., 2002).

The main task of the filter is to distinguish parallel document candidates from those that might have no parallel texts. In order to find out if Zipf's law is applicable to parallel documents, we analyzed the frequency distribution of a small parallel corpus. Table 1 shows the statistics of these data. We observed that our tiny-size corpus almost conform to the Zipf's law for the relationship of the rank and frequency of words in a corpus. Both the source and target languages show largely the same shape of relationship for the logarithm of rank and frequency. By analogy of the whole parallel corpus, we reached two linear functions for both languages with a slope close to -1. Figure 1 shows this observation.

The small size of corpora with this observation leads us to infer that this relationship should be held for two parallel documents as well. In two bilingual parallel documents, the rank and frequency of constituting words probably would be close to each other in two languages (The corresponding words in both sides should have largely the same rank and frequency). If two articles in two languages show the same pattern of relationship (a curve with the same slope) between the words ranks and frequencies, then we can infer that the two articles may have some degree of parallelism. In such cases, if a document in the source language consists of the words that have the ranks between 1 to r_s then the corresponding comparable document in the target language includes words ranks from 1 to r_t. Based on Zipf's law, r_s and r_t have a high probability to be close to each other. Intuitively, the area beneath the two functions as an indicator of parallelism of two documents would be close to each other. Figure 2 illustrates the idea where two candidate documents have some degree of parallelism versus two documents that are not related at all. We compute the area beneath the curve as *cumulative frequency log* for a document D as follows.



(a)

(b)

Figure 2: (a) two comparable documents that share parallel texts; (b) two documents that do not contain parallel texts.

$$Score(D) = \sum_{r=1}^{rmax} log(f(r)) \tag{2}$$

where r is the rank of the words in D, $rmax$ is the last rank in the document, and $f(r)$ is the frequency associated to the rank r.

Analyzing the cumulative frequency log of parallel documents reveals that for a given language, this score is linearly related to its counterpart in the other language. Figure 3 depicts this relationship for 40 Spanish-English parallel documents that are generated from Spanish part of Europarl corpora (Koehn, 2005). In this set, the lengths of document pairs are considered different.

Therefore, having the *Cumulative Frequency Log* of source documents will estimate the *Cumulative Frequency Log* of the target documents. In the training process with a set of n parallel documents, we use a Linear Regression Model to predict the response to n data points (x_1, y_1),(x_2, y_2), ...,(x_n,y_n) where x_i and y_i are the cumulative frequency log of ith parallel document pair in the source and target language, respectively. The linear regression model is given by

$$y = a_0 + a_1 x \tag{3}$$

where a_0 and a_1 are the constants of the regression model. A measure of best-fitting line, i.e, how well $a_0 + a_1 x$ predicts the cumulative frequency log of y is the magnitude of the error of predictions (ϵ_i) at each of the n data points.

$$\epsilon_i = y_i - (a_0 + a_1 x_i) \tag{4}$$

The regression parameters can be obtained by minimizing these errors of predictions by Least Square methods.

In the core of the filter, with two given documents in the source and target languages, namely D_s and D_t, the cumulative frequency log of two documents are computed as $x = Score(D_s)$ and $y = Score(D_t)$. Then x is put to the regression model to obtain the predicted cumulative frequency log of target document. By computing the absolute value of error of prediction (ϵ), we determine the parallelism of two documents if ϵ is smaller than or equal to a threshold called δ.

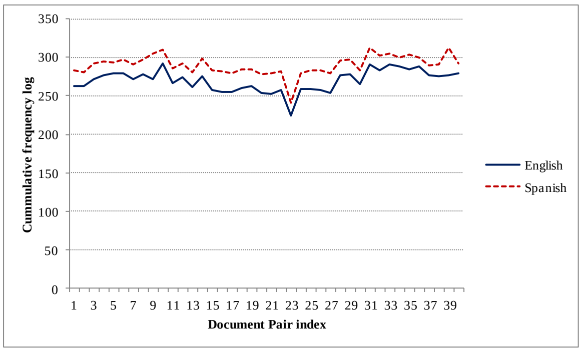


Figure 3: Cumulative frequency log of parallel documents

$$Par(D_s, D_t) = \begin{cases} 1, & \text{if } |\epsilon| \leq \delta \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

The best result for Eq. 5 is obtained when $\epsilon = 0$ which means the predicted value coincides with the actual value. However, we need to allow some degree of deviation from the regression model using δ. We can find the best value for δ that maximizes the precision and recall of the filter at the same time. Our experiments in the next section find different best δ for different language pairs.

4. Experiment and Results

We have used the English-Spanish (en-es), English-Dutch (en-nl), and English-Swedish (en-sv) parallel corpora in the Europarl dataset (Koehn, 2005) to evaluate our proposed method. In this regard, we split each parallel corpus to 77 parallel document pairs with different sizes. The range of size of these documents is from a couple of lines to about

Language pair	#test doc pairs	#parallel test docs	training data (MB)	
			source	target
English-Spanish (en-es)	314	27	182	201
English-Dutch (en-nl)	336	12	184	203
English-Swedish (en-sv)	290	26	170	177

Table 2: Statistical information of test and training dataset.

δ	English-Spanish			English-Dutch			English-Swedish		
	precision	**recall**	**F1**	**precision**	**recall**	**F1**	**precision**	**recall**	**F1**
1	0.57	0.15	0.24	0.57	0.33	0.42	0.86	0.23	0.36
2	0.60	0.22	0.32	0.47	0.58	0.52	0.69	0.35	0.46
3	0.62	0.30	0.40	0.47	0.75	0.58	0.74	0.65	0.69
4	0.55	0.41	0.47	0.41	0.75	0.53	0.71	0.77	0.74
5	0.52	0.44	0.48	0.41	0.92	0.56	0.65	0.77	0.70
6	0.50	0.67	0.57	0.37	0.92	0.52	0.62	0.81	0.70

Table 3: Evaluation results of the proposed method applied on three language pairs.

Length ratio threshold (β)	English-Spanish	English-Dutch	English-Swedish
0.1	0.74	0.44	0.57
0.2	0.47	0.31	0.47
0.3	0.36	0.21	0.37

Table 4: Accuracy of length-based filter to identify parallel documents

100K lines in which each line represents a sentence. For each language pair, we use 50 document pairs for training the model and use the remaining document pairs to create test data. The test data are generated using randomly picking one document from the source language and one from the target language. Actual parallel documents are identified by a same name in the source and target languages. Table 2 shows some statistical information about the training and test data.

In the experiment, we perform several runs with different threshold (δ) from 1 to 6. We go through interval of 1 for δ since we can see bigger changes in the precision and recall. Table 3 summarizes the precision, recall and F measure obtained by the proposed approach for three language pairs. Figure 4 illustrates the precision results for three given language pairs with varying δ. Figure 5 also shows the recalls with the same settings.

Our results show that using a low threshold yields higher precision and lower recall compared to using a high threshold that leads to lower precision and higher recall. We can rely on F-measure to find out the best setting for threshold. From the results in Table 3, the thresholds that maximize the F-measure for Spanish-English, Dutch-English, and Swedish-English are 6, 3, and 4, respectively. With these best configurations in the language pairs of the study, the filter achieves a precision between 0.47 to 0.71, recall between 0.67 to 0.77, and F-measure between 0.57 to 0.74.

Compared to the related works like (Zafarian et al., 2015) and (Morin et al., 2015) in which the precision is reported as 0.46 and 0.57, respectively, our approach achieves competitive results, in particular when the parameter δ is fine-tuned.

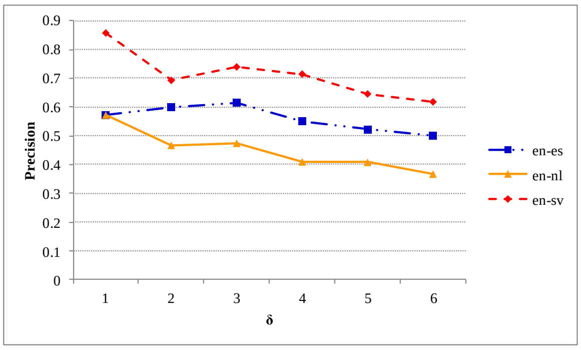


Figure 4: Precision trend versus delta for three language pairs.

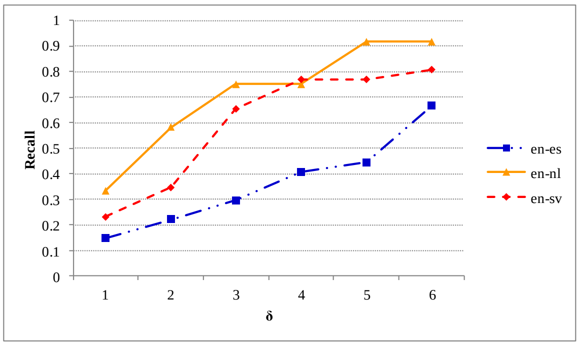


Figure 5: Recall of the proposed method with different delta for three language pairs.

We also run another experiment over test data using a length-based filter to identify parallel documents and benchmark against the proposed Zipfian-based filter. We compute the length ratio of each two documents i and j ($length\_ratio_{ij}$) based on their word counts and decide over their parallelism if $|length\_ratio_{ij} - 1| \leq \beta$, where β is a predefined threshold. Table 4 presents the precision results obtained by this method using different threshold values. The results show that the length based filter performs relatively well for English-Spanish documents, but its performance for English-Dutch and English-Swedish is not very good. In contrast, our Zipfian-based filter outperforms the length based filter for English-Dutch and English-Swedish documents.

5. Conclusion and Future Works

Parallel texts are an essential source of NLP and machine translation tasks while they are hardly available for under-resource languages. In this paper we proposed to identify parallel documents from a set of comparable articles using a filter based on Zipfian characteristic of parallel documents. We performed experiments over three language pairs to evaluate the proposed approach. Based on our results, the approach achieves promising results in terms of precision and recall of the identified parallel documents. The proposed method is language independent and does not rely on any linguistic knowledge.

Potential pathways for future works include extensive evaluation of the proposed method on larger experiment test cases that covers more language families. Another pathway would be to apply the proposed approach to some well-known existing methods for parallel text identification to improve the phase of document-level alignment in these approaches. In particular, applying the proposed method on linked Wikipedia articles to extract parallel articles from Wikipedia resources would be beneficial for low-resource languages.

6. References

Deane, P. (2005). A nonparametric method for extraction of candidate phrasal terms. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 605–613. Association for Computational Linguistics.

Ha, L. Q., Sicilia-Garcia, E. I., Ming, J., and Smith, F. J. (2002). Extension of zipf's law to words and phrases. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–6. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Malmasi, S., Refaee, E., and Dras, M. (2015). Arabic dialect identification using a parallel multidialectal corpus. In *Proceedings of the 14th Conference of the Pacific Association for Computational Linguistics (PACLING 2015), Bali, Indonesia*, pages 209–217.

Morin, E., Hazem, A., Boudin, F., and Clouet, E. L. (2015). Lina: Identifying comparable documents from wikipedia. In *Eighth Workshop on Building and Using Comparable Corpora*.

Pal, S., Pakray, P., and Naskar, S. K. (2014). Automatic building and using parallel resources for smt from comparable corpora. In *Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra)@ EACL*, pages 48–57.

Paramita, M. L., Guthrie, D., Kanoulas, E., Gaizauskas, R., Clough, P., and Sanderson, M. (2013). Methods for collection and evaluation of comparable documents. In *Building and Using Comparable Corpora*, pages 93–112. Springer.

Zafarian, A., Aghasadeghi, A., Azadi, F., Ghiasifard, S., Alipanahloo, Z., Bakhshaei, S., and Ziabary, S. M. M. (2015). Aut document alignment framework for bucc workshop shared task. *ACL-IJCNLP 2015*, page 79.

Exploring the Richness and Limitations of Web Sources for Comparable Corpus Research

Gregory Grefenstette
Inria Saclay/TAO, Rue Noetzlin - Bât 660
91190 Gif sur Yvette, France
gregory.grefenstette@inria.fr

Abstract

Comparable Corpora have been used to improve statistical machine translation, for augmenting linked open data, for finding terminology equivalents, and to create other linguistic resources for natural language processing and language learning applications. Recently, continuous vector space models, creating and exploiting word embeddings, have been gaining in popularity in more powerful solutions to creating, and sometimes replacing, these resources. Both classical comparable corpora solutions and vector space models require the presence of a large quantity of multilingual content. In this talk, we will discuss the breadth of this content on the internet to provide some type of intuition in how successful comparable corpus approaches will be in achieving its goals of providing multilingual and cross lingual resources. We examine current estimates of language presence and growth on the web, and of the availability of the type of resources needed to continue and extend comparable corpus research. .

Keywords: web mining, under-resourced languages, comparable corpora, language resources

Bibliographical References

Barbaresi, A. (2015). *Ad Hoc And General-Purpose Corpus Construction From Web Sources* (Doctoral dissertation, ENS Lyon).

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. Language resources and evaluation, 43(3), 209-226.

Dimitrova, L., Ide, N., Petkevic, V., Erjavec, T., Kaalep, H. J., & Tufis, D. (1998, August). Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central And Eastern European Languages. In *Proceedings Of The 17th International Conference On Computational Linguistics*, ACL, Volume 1 pp. 315-319.

Gatto, M. (2011). The 'Body' and The 'Web': The Web As Corpus Ten Years On. *ICAME J.*, 35, 35-58.

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *LREC* , pp. 759-765.

Grefenstette, G., & Nioche, J. (2000). Estimation of English and non-English Language Use on the WWW. In *Content-Based Multimedia Information Access-Volume 1*, RIAO 2000, pp. 237-246.

Hale, S. A. (2012). Net Increase? Cross-Lingual Linking in the Blogosphere. Journal of Computer-Mediated Communication, 17(2), 135-151.

Kilgarriff, A., & Grefenstette, G. (2003). Introduction to the Special Issue On The Web As Corpus. *Computational Linguistics*, 29(3), 333-347.

Pimienta, D., Prado, D., & Blanco, Á. (2009). Twelve Years Of Measuring Linguistic Diversity In *The Internet: Balance And Perspectives.* Paris: United Nations Educational, Scientific and Cultural Organization.

Rehm, G., & Uszkoreit, H. (2011). Multilingual Europe: A Challenge For Language Tech. *MultiLingual*, 22(3), pp. 51-52.

Ronen, S., Gonçalves, B., Hu, K. Z., Vespignani, A., Pinker, S., & Hidalgo, C. A. (2014). Links That Speak: The Global Language Network And Its Association With Global Fame. *Proceedings of the National Academy of Sciences*, 111(52), E5616-E5622.

Scannell, K. P. (2007). The Crúbadán Project: Corpus Building for Under-resourced Languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, Vol. 4, pp. 5-15.

Soria, C., Calzolari, N., Monachini, M., Quochi, V., Bel, N., Choukri, K., Mariani, J., Odijk, J. and Piperidis, S., (2014). The Language Resource Strategic Agenda: the FLaReNet Synthesis Of Community Recommendations. *Language Resources and Evaluation*, 48(4), pp.753-775.

Van der Veken, A., & De Schryver, G. M. (2003). Les Langues Africaines Sur La Toile: Étude Des Cas Haoussa, Somali, Lingala Et Isixhosa [The African Languages on the Internet: Case Studies for Hausa, Somali, Lingala and isiXhosa]. *Cahiers du RIFAL*, 23, 33-45.

A Mutual Iterative Enhancement Model for Simultaneous Comparable Corpora and Bilingual Lexicons Construction

Zede Zhu, Xinhua Zeng, Shouguo Zheng, Xiongwei Sun, Shaoqi Wang, Shizhuang Weng
Institute of Technology Innovation, Hefei Institutes of Physical Science, Chinese Academy of Sciences
Hefei Anhui, 230088, China
zhuzede@126.com, xhzeng@iim.ac.cn, zhshg1985@163.com, xiongweisun@gmail.com, wsq2012@mail.ustc.edu.cn, weng1989@mail.ustc.edu.cn

Abstract

Constructing bilingual lexicons from comparable corpora has been investigated in a two-stage process: building comparable corpora and mining bilingual lexicons, respectively. However, there are two potential challenges remaining, which are out-of-vocabulary words and different comparability degrees of corpora. To solve above problems, a novel iterative enhancement model is proposed for constructing comparable corpora and bilingual lexicons simultaneously under the assumption that both processes can be mutually reinforced. As compared to separate process, it is concluded that both simultaneous processes show better performance on different domain data sets via a small-volume general bilingual seed dictionary.

Keywords: comparable corpora, bilingual lexicons, mutual iterative enhancement, simultaneous construction

1. Introduction

Comparable corpora are selected as pairs of mono-lingual documents based on the criteria of content similarity, non-direct translation and language difference. With respect to parallel corpora, comparable corpora have the advantages in terms of more up-to-date, abundant and accessible (Ji et al., 2009). Furthermore, they are valuable resources for multilingual information processing, from which parallel sentences (Smith et al., 2010), parallel phrases (Munteanu and Marcu, 2006) and bilingual lexicons (Li and Gaussier, 2010; Prochasson and Fung, 2011) can be mined to reduce the sparseness of existing resources (Munteanu and Marcu, 2005; Snover et al., 2008).

Note that previous works of bilingual lexicons construction from comparable corpora consist of two stages separately: building comparable corpora and mining bilingual lexicons (Figure 1(a)). In the first stage, the automatic building of comparable corpora can be completed by focused crawling, cross-language information retrieval or 'inter-wiki' link. However, utilizing the comparability degree to build comparable corpora is still a significant challenging task. The degree of comparability is usually defined as the expectation of finding the translation of source language vocabularies in the target language documents. Therefore, most methods adopt statistical approach to map vocabularies in different languages by a bilingual seed dictionary.

In the second stage, the seminal works of mining bilingual lexicons from comparable corpora are based on the word co-occurrence hypothesis, in which the word and its translation share similar contexts. They assume the corpora are reliably comparable and focus on the improvement of extraction algorithms (Hazem et al., 2012), whereas successful detection of bilingual lexicons is severely influenced by the quality of corpora.

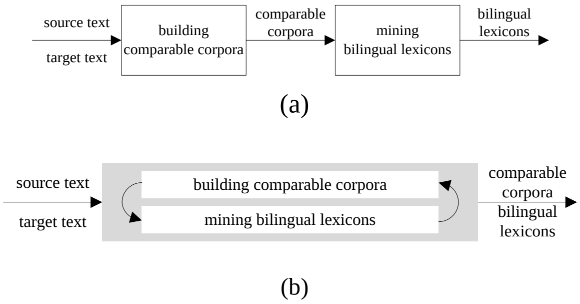


Figure 1: (a) Separate comparable corpora construction and bilingual lexicons construction and (b) joint comparable corpora construction and bilingual lexicons construction.

These two stages respectively suffer from different major challenges: Firstly, if the seed dictionary and the document set are less relevant in domain, out-of-vocabulary words will lower the quality of comparable corpora, especially domain-specific words. Secondly, if the comparable corpora have low comparability degrees, the quality of corpora may limit the performance of the bilingual lexicons construction. To address these potential problems, a novel iterative enhancement model is proposed to construct comparable corpora and bilingual lexicons simultaneously under the assumption that both processes can be mutually boosted (Figure 1(b)). The similar model has success in the domain of cross-domain sentiment classification (Wu et al., 2010).

Contributions Our contributions are as follows:

① A novel iterative enhancement model is presented to construct two different grained size levels bilingual resources simultaneously.

② A novel method of enriching domain-specific bilingual lexicons directly harvested from the candidate comparable corpora is proposed to enhance the ability of building comparable corpora.

③ A novel method of calculating the relativity of cross-language lexicons on the basis of different comparability degrees of comparable corpora is proposed to enhance the ability of mining bilingual lexicons.

④ The model can be effectively applied in various domains, even though it relies on fewer existing resources such as a small-volume general bilingual seed dictionary.

The research hypothesis and motivation are just presented in this section. In the following section, we briefly summarize state-of-the-art approaches of the comparable corpora and bilingual lexicons construction. The iterative enhancement algorithm is described in detail in the "Proposed Model" section. Finally, we present our datasets, experiments and results before concluding the paper.

2. Related Work

2.1 Comparable Corpora Construction

The automatic acquisition of multilingual corpora can be completed by a variety of methods: focused crawling (Talvensaari et al., 2008), cross-language information retrieval (Huang et al., 2010) and 'intewiki' links (Smith, 2010). In fact, the measuring comparability degree of document pairs is still a challenging task to construct comparable corpora.

Recent measuring works mainly adopt statistical approach to map common vocabularies in different languages. To map lexical items, (Li and Gaussier, 2010) made use of a translation table and (Su and Babych, 2012) adopted a bilingual seed dictionary. (Saad et al., 2013) proposed two different comparability measures based on binary and cosines similarity measures using the bilingual dictionary to align words. Given a comparable corpora, (Li and Gaussier, 2010; Su and Babych, 2012) defined the degree of comparability as the expectation of finding the translation of any given source/target words in the target/source corpora vocabulary. In addition (Zhu et al, 2013) utilized the trained bilingual LDA model to calculate the comparability.

These approaches effectively evaluate the metric on the rich-resourced language pairs, thus quality bilingual resources are available. However, this is not the case for all domains in which reliable language resources such as bilingual dictionaries with broad word coverage might be not publicly available. To avoid the limit of existing resources, Tao and Zhai (2005) proposed a purely language-independent method to extract comparable bilingual text without the existing linguistic resources. They assumed that two words with mutual translation should have similar frequency correlation. The association between two documents was then calculated based on this information.

Nevertheless, the performance of the above method may be compromised due to the lack of linguistic knowledge, particularly corpora with low comparability. In this article, the problem can be circumvented by enriching a small general bilingual seed dictionary with a domain-specific bilingual lexicons harvested gradually from candidate comparable corpora to increase the dictionary coverage facing source and target texts.

2.2 Bilingual Lexicons Construction

The seminal works of extracting bilingual lexicons from comparable corpora are based on the word co-occurrence hypothesis, where the term and its translation share similar contexts (Fung, 1998; Rapp, 1999). More recent works usually assume that corpora are reliably comparable and focus on the improvement of extraction algorithms (Hazem et al., 2012). Therefore, less work is focused on the characteristics of comparable corpora (Maia, 2003). In fact, the degree of comparability has the greatly divergence between different corpora. Usually, successful detection of bilingual lexicons from comparable corpora depends on the quality of corpora, especially the degree of their textual equivalence and successful alignment on various text units.

To extract high-quality lexicons, the target and source texts should be highly comparable in a very specific subject domain. If one arbitrarily increases the size of the corpora, he actually takes the risk of decreasing its quality by adding out-of-domain texts. It has been proved that the quality of the corpora is more important than its size. Morin et al. (2007) showed that the discourse categorization of the documents increases the precision of the lexicons despite of the data sparsity. (Li and Gaussier, 2010; Li and Gaussier, 2011) improved the quality of the extracted lexicons when they improved the comparability of the corpora by selecting a smaller–but more comparable corpora from an initial set of documents. (Su and Babych, 2012) presented three different approaches to measure the comparability of cross-lingual comparable documents: a lexical mapping, a keyword and a machine translation approach. The results proved that higher comparability level consistently resulted in more number of parallel phrases extracted from comparable documents. Moreover, (Wang el at., 2014) adopted two step cross-comparisons between translation candidates to improve the quality.

Nevertheless, these methods couldn't effectively make use of comparable corpora of low comparability degree discarded directly. In this article, according to characterize the different comparability, the candidate comparable corpora is awarded different weight to extract good-quality bilingual lexicons from the corpora along with traditional context information.

3. Proposed Model

3.1 Basic Concepts Representation

The model is based on the assumption that the comparability of document pairs can promote the similarity of word pairs, and the similarity of word pairs can enhance the comparability of document pairs, which completes a mutual iterative enhancement model for simultaneous comparable corpora and bilingual lexicons construction shown in Figure 2.

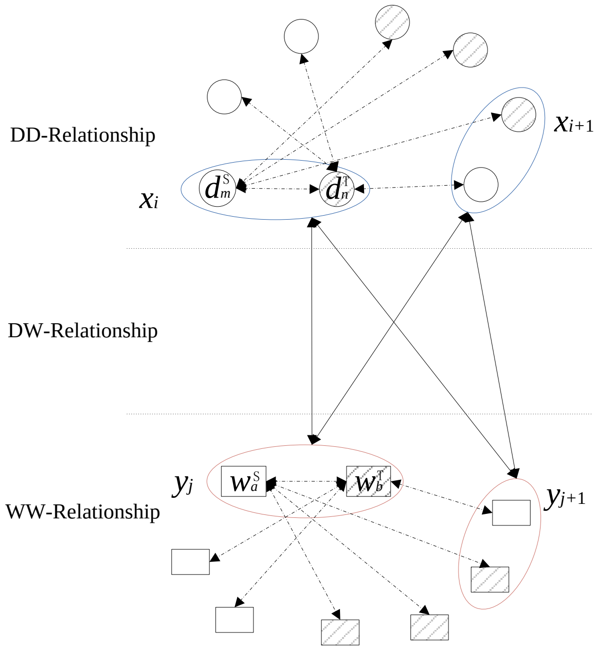


Figure 2: An illustration of the joint model between documents and words, where ○ (d_m^S) means source language document, ⊘ (d_n^T) means target language document, □ (w_a^S) means source language word, ▨ (w_b^T) means target language word, ⬭ (x_i) describes bilingual document pairs consisting of d_m^S and d_n^T, and ⬭ (y_j) describes bilingual word pairs consisting of w_a^S and w_b^T.

3.2 Relationship Formation

To establish the several relationships, the two basic functions have been proposed to measure:

$$\sigma(w^S, w^T) = \begin{cases} 1, if\ w^S\ is\ a \text{ translation of } w^T \\ 0, otherwise \end{cases} \quad (1)$$

Where the function $\sigma(w^S, w^T)$ checks whether the translation of the word w^S in the source language document is equal to another word w^T in its corresponding target language document.

$$\delta(\mathrm{w}, w') = \begin{cases} 1, if\ w\ and\ w'\ are\ equivalents \\ 0, otherwise \end{cases} \quad (2)$$

Where the function $\delta(w, w')$ checks whether two words w and w' are equivalence in the same language document.

Step 1: Calculating DD-Relationship

Given the source language document collection $D_S = \{d_m^S | 1 \leq m \leq \mathrm{M}\}$ (M represents the number of source language documents) and the target language document collection $D_T = \{d_n^T | 1 \leq n \leq \mathrm{N}\}$ (N represents the number of target language documents), the comparability degree R_{x_i} between multilingual document pairs x_i can be defined by the rate of translation between d_m^S and d_n^T, which is calculated by two bilingual unidirectional seed dictionaries. R_{x_i} is produced by the following formula:

$$R_{x_i} = P_{sim}(d_m^S, d_n^T) \quad (3)$$

$$= \frac{\#_w^{trans}(d_m^S, d_n^T)}{\#_w d_m^S} + \frac{\#_w^{trans}(d_n^T, d_m^S)}{\#_w d_n^T}$$

$$= \frac{1}{\#_w d_m^S} \sum_{w_e^S \in d_m^S} \sum_{w_f^T \in d_n^T} \sigma(w_e^S, w_f^T)$$

$$+ \frac{1}{\#_w d_n^T} \sum_{w_f^T \in d_n^T} \sum_{w_e^S \in d_m^S} \sigma(w_f^T, w_e^S)$$

Where $\#_w *$ is the number of words in the document *; $\#_w^{trans}(d_m^S, d_n^T)$ is the number of translation glossaries from d_m^S to d_n^T in different languages. If the comparability degree R_{x_i} exceeds the predefined threshold R_0, the cross-lingual document pair x_i forms an initial candidate multilingual comparable document pair whose weight is recorded as R_{x_i}.

Step 2: Calculating WW-Relationship

Given the source language word collection $W_S = \{w_a^S | 1 \leq a \leq \mathrm{A}\}$ (A represents the number of source language words) and the target language word collection $W_T = \{w_b^T | 1 \leq b \leq \mathrm{B}\}$ (B represents the number of target language words), the statistical relationship L_{y_j} between two words w_a^S and w_b^T can be calculated by the mutual information on the basis of the co-occurrence information. L_{y_j} between w_a^S and w_b^T is calculated as follows:

$$L_{y_j} = P_{co}(w_a^S, w_b^T) = \frac{2 \times \#(w_a^S, w_b^T)}{\#(w_a^S) \times \#(w_b^T)} \quad (4)$$

$$= \frac{2}{\left[\sum_{w_g^S \in D_S} \delta(w_a^S, w_g^S)\right] \times \left[\sum_{w_p^T \in D_T} \delta(w_b^T, w_p^T)\right]}$$

$$\times \sum_{x_i \in \mathrm{X}} min \left\{ \sum_{w_g^S \in d_m^S} \delta(w_a^S, w_g^S), \sum_{w_p^T \in d_n^T} \delta(w_b^T, w_p^T) \right\}$$

Which indicates the degree of statistical dependence between w_a^S and w_b^T. Here, $\#(w_a^S, w_b^T)$ is the number of w_a^S and w_b^T co-occurrence in all candidate comparable document pairs; $\#(w_a^S)$ and $\#(w_b^T)$ are respectively the frequencies of w_a^S and w_b^T in the document collection. If L_{y_j} exceeds the predefined threshold L_0, a cross-lingual word pair y_j is considered as an initial candidate bilingual lexicons pair whose weight is marked as L_{y_j}.

Step 3: Calculating DW-Relationship

Given the candidate bilingual document pairs collection $X = \{x_i | 1 \leq i \leq \mathrm{I}\}$ (I represents the number of the candidate bilingual document pairs) and the candidate bilingual word pairs collection $Y = \{y_j | 1 \leq j \leq \mathrm{J}\}$ (J represents the number of the candidate bilingual word pairs), a weighted bipartite relationship $H_{x_i y_j}$ between x_i and y_j can be calculated by the following formula when the word pair y_j appears in the document pair x_i.

$$H_{x_iy_j} = P_{rel}\left(R_{x_i}, L_{y_j}\right) = \frac{2 \times \#_{2w}^{co}\left(x_i, y_j\right)}{\#_w d_m^S + \#_w d_n^T} \qquad (5)$$

$$= \frac{2}{\#_w d_m^S + \#_w d_n^T}$$

$$\times\, min\left\{ \sum_{w_h^S \in d_m^S} \delta\left(w_a^S, w_h^S\right), \sum_{w_r^T \in d_n^T} \delta\left(w_b^T, w_r^T\right) \right\}$$

Where $\#_{2w}^{co}(x_i, y_j)$ is the number of the times that w_a^S and w_b^T co-occur in the document pair x_i. $H_{x_iy_j}$ can indicate the degree of statistical dependence between x_i and y_j.

3.3 Iterative Enhancement Algorithm

The core of the algorithm is to calculate the reasonable values of variables R_{x_i} and L_{y_j}. When the algorithm is carried out in the t^{th} iteration, the R_{x_i} and L_{y_j} are denoted as the $R_{x_i}^t$ and $L_{y_j}^t$ respectively. In order to calculate the values of $R_{x_i}^t$ and $L_{y_j}^t$, the iterative enhancement algorithm is mainly proposed on the basis of two basic assumptions as follows:

① If each document pair x_i in different languages contains more bilingual translation vocabularies, x_i should have a greater likelihood to construct comparable corpus;

② If each word pair y_j in different languages appears in the comparable corpora with high comparability degree, y_j should have a greater likelihood to construct bilingual lexicon.

According to the above assumptions, the change of $R_{x_i}^t$ is mainly dependent on $L_{y_j}^{t-1}$, and the change of $L_{y_j}^t$ is mainly dependent on $R_{x_i}^{t-1}$, where the initial values $R_{x_i}^0$ and $L_{y_j}^0$ respectively are calculating with formulas (3) and (4). When $R_{x_i}^t$ is greater than a predefined threshold R, x_i is a candidate comparable corpus. When $L_{y_j}^t$ is greater than a predefined threshold L, y_j is a candidate bilingual word pair. Finally, we can establish the following iterative forms:

$$R_{x_i}^t = \alpha R_{x_i}^{t-1} + \beta \sum_{j=1, L_{y_j}^{t-1} > L}^{J} H_{x_iy_j} L_{y_j}^{t-1} \qquad (6)$$

$$L_{y_j}^t = \alpha L_{y_j}^{t-1} \qquad (7)$$

$$+\beta \sum_{i=1, R_{x_i}^{t-1} > R}^{I} \left(H_{y_jx_i} + \cos < \vec{C}_{w_a^S}, \vec{C}_{w_b^T} >\right) R_{x_i}^{t-1}$$

Where $\alpha + \beta = 1$, α and β specify the relative contributions to the final scores; The value of $H_{y_jx_i}$ is equal to the value of $H_{x_iy_j}$, which remains unchanged in the iterative process. $\vec{C}_{w_a^S}$ and $\vec{C}_{w_b^T}$ are respectively the context vectors of w_a^S and w_b^T. $\cos < \vec{C}_{w_a^S}, \vec{C}_{w_b^T} >$ is calculated by the standard approach (Fung, 1998; Rapp, 1999).

Finally, the convergence of the iteration algorithm is achieved when the difference of every document pair and word pair falls below a predefined threshold θ, which is formally expressed by the following two formulas: $\left|R_{x_i}^t - R_{x_i}^{t-1}\right| < \theta$ and $\left|L_{y_j}^t - L_{y_j}^{t-1}\right| < \theta$.

4. Experiments and analysis

In this section, several experiments are conducted to verify the effectiveness of this model. The initial thresholds are set as follows: $L_0 = 0.15$, $R_0 = 0.3$, $L = 0.1$, $R = 0.1$ and $\theta = 0.0001$, which are identified by the previous works.

Questions We try to answer the following questions:

① Does the joint model outperform conventional methods of building comparable corpora? (Section 4.1)

② How about the quality of lexicons by the joint model of mining bilingual lexicons? (Section 4.2)

4.1 Comparable Corpora Evaluation

4.1.1. Evaluation Measures

As there is no commonly available data set to evaluate the comparability degree of comparable corpora and then mine bilingual lexicons, we collect our own gold standard comparable corpora as test datasets. They specialize on three different domains on *culture*, *economy* and *sport*, which include 50 English-Chinese bilingual document pairs respectively. The datasets are normalized through the following linguistic preprocessing steps: tokenization, part-of-speech tagging, lemmatization and function word removal. In addition, a small-volume general bilingual seed dictionary is applied which contains 42,373 distinct common entries.

The datasets are acquired by two main steps. Firstly, the initial data are acquired by adopting the focused crawling for automatic acquisition of topic-specific source language web and utilizing interlinks between pages to collect target language web. This method can quickly locate a relative specific domain including 500 page pairs. Secondly, we manually annotate the document pairs on the basis of five comparability levels as gold standard to assess the alignments. Five levels proposed by (Fung P. 1998) are refined the alignments as follows: Same Story, Related Story, Shared Aspect, Common Terminology and Unrelated. Finally, we select 50 document pairs in every domain with Same Story and Related Story as comparable corpora.

We adopt the *Precision* as evaluation metric:

$$Precision = \left|C_p \cap C_l\right| / \left|C_p\right| \qquad (8)$$

Where C_p represent the comparable corpora in the automatic building results; C_l represent the comparable corpora in the labeled results; $|*|$ means the number of document pairs in the corpora $*$.

4.1.2. Results and Analysis

We set two parameters α=0.5 and β=0.5 according to the conclusion of the 'Group 1' in the 4.2.2 subsection. Then we compare the performance of the joint model with the current representative approach (shown in Table 1).

domain		*culture*	*economy*	*sport*
This paper	No-iterative	45	57	49
	Iterative	64	83	69
	Value of improvement	↑ 19	↑ 26	↑ 20
Zhu et al. (2013)		58	77	67

Table 1: Performance (%) of the *Precision* for different domains and existing method.

Overall, the results indicate the robustness and effectiveness of the model. It is concluded that the model can be effectively applied to different domains even through external resources is under adverse conditions that the seed dictionary is a small-volume general bilingual dictionary. In every specific domain, the results reliably depend on the correlation of cross-language document pairs in the datasets. Simultaneously, with respect to the no-iterative process, the performance of the iterative enhancement significantly improves up to 26%. In addition, the scores of this paper outperform the algorithm implemented by (Zhu et al, 2013), which adopts the trained bilingual LDA model to predict the topical structures and calculates the similarity of the documents in different languages. The high quality results of the joint model are due to the fact that out-of-vocabulary words are sufficiently solved in this paper.

4.2 Bilingual Lexicons Evaluation

4.2.1. Evaluation Measures

Automatic evaluation of bilingual lexicons extraction is performed against a gold standard lexicons *G*, which is obtained from the top-ranking nouns or verbs in the gold standard comparable corpora. These lexical items should only appear in a domain bilingual dictionary and be not included in the seed dictionary that is a small-volume general bilingual dictionary. *G* contains 100 Chinese single-word terms with their corresponding English translations. When more than one translation variant are possible for a single English term, each proposed by the model is considered as correct result.

We adopt the *Accuracy* as evaluation metric in bilingual lexicons extraction, which reflects precision among first *K* translation candidates. And the *Accuracy* is calculated in the following equation:

$$Accuray = count_{top\ K}/H \qquad (9)$$

Where *H* means the number of the gold standard entries in *G*; $count_{top\ K}$ means all the number of correct translation in top *K* ranking. In this paper, *K* ranges from 1^{th} to 20^{th} ranking.

4.2.2. Results and Analysis

Group 1: Parameter β

In order to better grip the relative contributions from the document x_i and the word y_j, table 2 shows the score with respect to the parameter β in the entire corpora collection and β ranges from 0 to 1 by 0.1 as a step length.

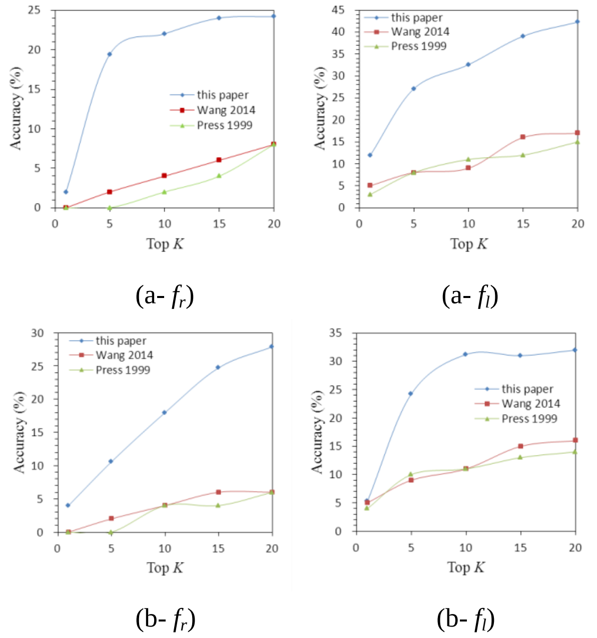
K\β	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
1	0	2	4	6	7	8	7	8	6	3	3
5	1	7	9	16	17	18	19	17	15	11	8
10	2	9	15	17	19	20	21	21	19	14	10
15	2	11	16	19	23	21	20	23	18	15	12
20	4	12	17	20	23	22	22	23	21	15	13
Total	9	41	61	78	89	**91**	90	90	79	58	46

Table 2: Performance (%) of the *Accuray* with value of varied *K* from 1 to 20 by 5 as a step length.

The table 2 shows that the parameter β has a remarkable impact on the performance of the model. When the value of β is set as 0.4 or 0.6, the *Accuracy* mostly achieves the peak with each value of *K*. If β becomes large enough (near to 1) or very small (near to 0), the *Accuracy* sharply falls into decline. These results demonstrate that both documents and words are very important contributions to rank comparable corpora. The loss of each element will greatly deteriorate the final performance. The total of *Accuracy*, which shows the overall performance of the algorithm with all values of *K*, arrives the best performance under the condition of β =0.5. So the optimal β is set to 0.5 in the subsequent experiments according to the analysis of influence.

Group 2: Existing Methods Comparison

In order to verify the excellence of the model in the paper, we make use of all the document pairs as test dataset. Then we compare the performance of our model with the other two existing representative approaches: one is proposed by (Press, 1999) which reflects a baseline level, the other one is proposed by (Wang el at., 2014) which represents the current state of the art (Shown in Figure 3).

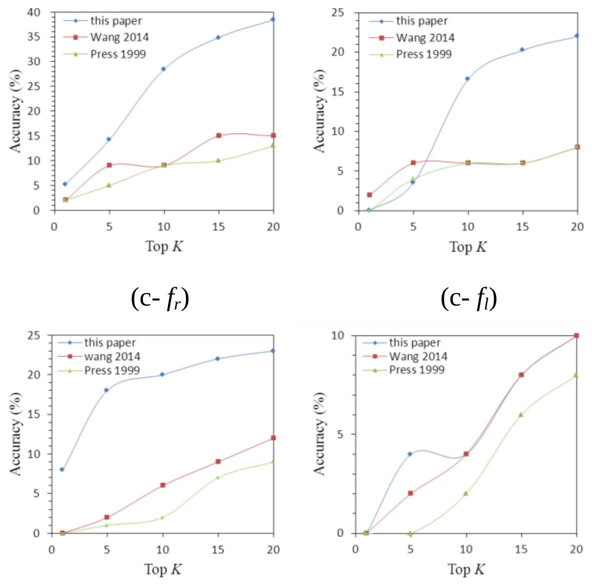


(a- f_r)

(a- f_l)

(b- f_r)

(b- f_l)



(c- f_r) (c- f_l)

(d- f_r) (d- f_l)

Figure 3: Performance of bilingual lexicons construction from different the methods with varied *K* values from 1 to 20. (a) culture, (b) economy, (c) sport, (d) mixed: containing three domains. (f_r) random frequency words, (f_l) low frequency words.

The figure 3 shows that the score obtained by this paper practically outperforms the other two approaches in three different domains regardless of word frequency, which indicates that iterative enhancement model is valid to construct bilingual lexicons. (Press, 1998) extracts bilingual lexicons in the view of the context information. (Wang el at., 2014) adopts two step cross-comparisons between translation candidates of each target word to improve the quality of bilingual lexicons. But the correlation between vocabularies completely depends on the coverage of seed dictionary and the comparability of the document pairs are utilized as equalization. When the dictionary cannot cover the most of glossaries in the corpora due to different domains, the method will loss the advantage.

The model proposed in the paper not only can distinguish the comparability degree of different document pairs to mine bilingual lexicons, but also utilize domain-specific bilingual lexicons producing in this process to calculate the comparability degree, which are continuous iteration and mutually reinforced. Only when low frequency bilingual lexicons are extracted from the mixed corpora, does the model proposed by this paper have almost equivalent performance with the method put forward by (Wang el at., 2014) shown in figure 3 (d- f_l). The main reason is that the mixed corpora have great differences of the domain knowledge, which lead to a very small promotion in the iterative process, especially when the target bilingual lexicons are the low frequency vocabularies.

5. Conclusions

Previous works on bilingual lexicons construction from comparable corpora are completed by two independent tasks. In this paper, we propose a simultaneous comparable corpora and bilingual dictionary construction method based on a mutual iterative enhancement model. Our evaluation shows the simultaneous construction approach improves the accuracy of the outcome comparable corpora and bilingual dictionary via a small-volume general bilingual seed dictionary. In addition, based on the encouraging results, we are going to explore more other sizes of bilingual resources simultaneously, such as bilingual parallel sentences and bilingual multi-word expressions.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of Anhui No.1608085QF127 and No. 1508085QC65, the Open Projects Program of National Laboratory of Pattern Recognition No.201306320, the National Natural Science Foundation of China No.31401285 and No. 61475163, China Postdoctoral Science Foundation No.2015M570548.

7. References

Ji H. (2009). Mining name translations from comparable cor-pora by creating bilingual information networks. In *Proceedings of BUCC*, pp. 34--37.

Jason Smith, Chris Quirk and Kristina Toutanova. (2010). Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment. In *Proceedings of NAACL*, pp. 403--411.

Dragos Munteanu and Daniel Marcu. (2006). Extracting parallel sub-sentential fragments from nonparallel corpora. In *Proceedings of ACL*, pp. 81--88.

Li B., and Gaussier E. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 644--652.

Emmanuel Prochasson and Pascale Fung. (2011). Rare Word Translation Extraction from Aligned Comparable Documents. In *Proceedings of ACL-HLT*, pp. 1327--1335.

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

Matthew G. Snover, Bonnie J. Dorr, and Richard M. Schwartz. (2008). Language and translation model adaptation using comparable corpora. In *Proceedings of EMNLP*, pp. 857--866.

Hazem A. and Morin E. (2012). Adaptive dictionary for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pp. 288--292.

Wu Q, Tan S, Cheng X, et al. (2010). MIEA: a mutual iterative enhancement approach for cross-domain sentiment classification. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pp. 1327--1335.

Tuomas Talvensaari, Ari Pirkola, Kalervo Järvelin, Martti Juhola, and Jorma Laurikkala. (2008). Focused web crawling in the acquisition of comparable corpora. *Information Retrieval*, 11(5), pp. 427--445.

Huang D. G., Zhao L., Li L. S., et al. (2010). Mining Large-scale Comparable Corpora from Chinese-English News Collections. In *Proceedings of Coling*, pp. 472--480.

Su F., Babych B. (2012). Measuring Comparability of Documents in Non-Parallel Corpora for Efficient Extraction of (Semi-) Parallel Translation Equivalents. In *Proceedings of the EACL*, pp. 10--19.

Motaz Saad, David Langlois, and Kamel Smaïli. (2013). Extracting comparable articles from wikipedia and measuring their comparabilities. *Procedia-Social and Behavioral Sciences*, 95(4), pp. 40--47.

Zhu Z., Li M., et al. (2013). Building Comparable Corpora Based on Bilingual LDA Model. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 278--282.

Tao Tao, Chengxiang Zhai. (2005). Mining comparable bilingual text corpora for cross-language information integration. In *Proceedings of ACM SIGKDD*, pp. 691--696.

Fung Press. (1998). A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora. *Machine Translation and the Information Soup*, pp. 1--17.

Rapp R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, pp. 519--526.

Maia B. (2003). Some languages are more equal than others. Training translators in terminology and information retrieval using comparable and parallel corpora. *Corpora in Translator Education*, pp. 43--53.

Morin E., Daille B., Takeuchi K., et al. (2007). Bilingual terminology mining-using brain, not brawn comparable corpora. In *Proceedings of the annual meeting of the Association for Computational Linguistics*, 45(1): 664--671.

Li B., Gaussier E., Aizawa A. (2011). Clustering comparable corpora for bilingual lexicon extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: short papers-Volume 2*, pp. 473--478.

Wang Shaoqi, Li Miao; et al. (2014). Improvement of Bilingual Lexicon Extraction Performance From Comparable Corpora via Optimizing Translation Candidate Lists. In *Proceedings of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pp. 18--25.

Hard Synonymy and Applications in Automatic Detection of Synonyms and Machine Translation

Ana Sabina Uban
Faculty of Mathematics and Computer Science, University of Bucharest
ana.uban@gmail.com

Abstract

We investigate in this paper the property of *hard synonymy*, defined as synonymy which is maintained across two or more languages. We use synonym dictionaries for four languages, as well as parallel corpora, and tools for distributional synonym extraction, in order to perform experiments to investigate the potential applications of hard synonymy for the automatic detection of synonyms and for machine translation. We show that hard synonymy can be used to discriminate between distributionally similar words that are true synonyms and those that are merely semantically related or even antonyms. We also investigate whether hard synonym word-translation pairs can be useful for lexical machine translation, by analyzing their occurrences in word-aligned parallel corpora. We build a database of words, synonyms and their translations for the four languages, including a generally low resourced language (Romanian) and show how it can be used to investigate properties of words and their synonyms cross-lingually.

Keywords: hard synonymy, Romanian, database, cross-lingual synonyms, distributional synonyms

1. Introduction

Synonymy is a lexical semantic relation, that is, a relation between meanings of words. By definition, synonyms are ‘words or expressions of the same language that have the same or nearly the same meaning in some or all senses’ (Merriam-Webster, 2004). Cross-linguistically, the question that we try to answer in this paper is how much of this common meaning is shared by pairs of translated words. Since synonymy closely associates different lexicalizations of the same concept (which is language-specific), the overlap between synonym sets across a pair of languages expresses a kind of concept lexicalization overlap.

Cross-lingual synonym sets prove to be useful in tasks such as, for instance, automatic translation of web pages. Since search engines are using more of the Latent Semantic Indexing, which associates keywords of an article or a page with its synonyms within the domain covered by the keywords, one needs to take into consideration the synonym set of the translated keywords and the overlap of two languages synonym sets.

2. Related Works

There are various NLP applications using synonyms, one of the most notable being automatic synonym detection or extraction (Wang and Hirst, 2011; Wang et al., 2010; Mohammad and Hirst, 2006; Bikel and Castelli, 2008), which in turn can help in tasks including machine translation, information retrieval, speech recognition, spelling correction, or text categorization (Budanitsky and Hirst, 2006).

A multilingual approach based on word alignment of parallel corpora proved to have (Van der Plas et al., 2011) higher precision and recall scores for the task of synonym extraction than the monolingual approach. Other work on semantic distance between words and concepts (Mohammad et al., 2007) emphasise on the advantages of multilingual over the monolingual treatment.

3. Hard Synonymy

Hard Synonymy is defined in (Dinu et al., 2015) as the semantic relation between two words that are synonyms in more than one language. In addition to their results, we add Spanish to the set of languages analyzed, and we provide a database containing all words in the four languages as found in synonym dictionaries, as well as their synonyms and their translations, and show how it can be used to extract hard synonyms. We then analyze the frequency and behavior of the synonym sets and word-translation pairs with this property and investigate their applications to synonym detection and to machine translation.

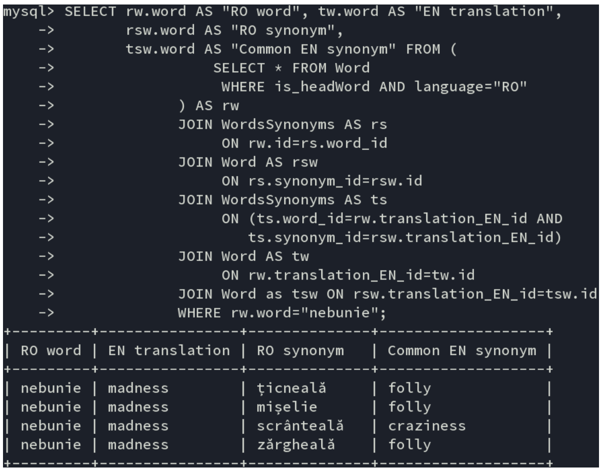
3.1. Resources

In order to obtain sets of hard synonyms we created a database with words from four different languages: English, French, Romanian and Spanish, along with their translations, and their synonyms. We used Google Translate API in order to translate every word into each of the other three languages, and synonym dictionaries for obtaining their synonyms. For English we employed Princeton’s WordNet, version 3.0; for French we used the synonyms dictionary developed by the CRISCO research centre; for Romanian we used a synonym dictionary (*Dicționarul de sinonime al limbii Române*, by Luiza Seche and Mircea Seche); and for Spanish we used Open Multilingual WordNet.

We organized the data in a MySQL database, in order to gain ease of access and to be able to instantiate various queries. The database consists of two tables: the first is the *Word* table - containing all words, as well as information on their translations, language and part of speech. There is a uniqueness constraint on the pair of columns (word, language), reflecting the uniqueness of word forms in each language. The second table is *WordsSynonyms* - containing synonymy relations as references to pairs of words in the *Word* table.

This database structure straightforwardly allows for queries such as, for instance, queries on synonym set overlap, function of the word pair’s part of speech tag.

An example of such a query, that extracts the common synonyms for the Romanian-English word pair *nebunie* - *madness*, is depicted in Figure 1 below.



```
mysql> SELECT rw.word AS "RO word", tw.word AS "EN translation",
    ->        rsw.word AS "RO synonym",
    ->        tsw.word AS "Common EN synonym" FROM (
    ->                  SELECT * FROM Word
    ->                   WHERE is_headWord AND language="RO"
    ->              ) AS rw
    ->              JOIN WordsSynonyms AS rs
    ->                   ON rw.id=rs.word_id
    ->              JOIN Word AS rsw
    ->                   ON rs.synonym_id=rsw.id
    ->              JOIN WordsSynonyms AS ts
    ->                   ON (ts.word_id=rw.translation_EN_id AND
    ->                      ts.synonym_id=rsw.translation_EN_id)
    ->              JOIN Word AS tw
    ->                   ON rw.translation_EN_id=tw.id
    ->              JOIN Word as tsw ON rsw.translation_EN_id=tsw.id
    ->              WHERE rw.word="nebunie";
+---------+----------------+-------------+-------------------+
| RO word | EN translation | RO synonym  | Common EN synonym |
+---------+----------------+-------------+-------------------+
| nebunie | madness        | ţicneală    | folly             |
| nebunie | madness        | mişelie     | folly             |
| nebunie | madness        | scrânteală  | craziness         |
| nebunie | madness        | zărgheală   | folly             |
+---------+----------------+-------------+-------------------+
```

Figure 1: An example of a database query

3.2. Methodology

In the pre-processing step, we extracted and cleaned the data in the Romanian and French dictionary, and removed multiword expressions for all languages. For further analysis we only consider the words for which translations were available using the Google Translate API; the number of such words for each language is illustrated in Table 1 below.

	Words	Translation pairs			
		EN	FR	RO	ES
EN	44.913	-	25.229	19.499	11.029
FR	40.765	22.338	-	20.789	11.011
RO	42.278	21.402	23.946	-	11.292
ES	10.028	7.942	8.070	7.062	-

Table 1: Number of words and translation pairs

Synonymy was considered a symmetric property - that is, for each *(w, s)* word-synonym pair found in the dictionaries, *(s, w)* was added as a synonym pair as well. Translation was generally not considered symmetric, but back-translations were used to fill in missing data where translations for some words in certain languages were not found by the API. In the case of homonyms or polysemantic words, we merged all the synonyms for each sense of the word together, thus obtaining unique word forms across the entire word set (for either of the four languages), each associated with one synonym set.

For each pair of languages among the four languages analyzed, we generated word-translation pairs, we then computed statistics on their respective synonym sets, measuring overlaps between sets of synonyms from two perspectives: first translating the original word's synonyms in order to find their overlap with the translation's synonyms, and then translating the translation's synonyms in order to find their overlap with the original word's synonyms, resulting in overlap scores for each language pair.

We also counted the number of word-translation pairs for which at least one common synonym was found, or the synonym overlap contained at least one synonym. The synonym sets that overlap across two languages will be called *hard synonyms*, and their corresponding word-translation pairs - *hard synonym pairs*.

3.3. Results: Hard Synonym Pairs

The percentage of hard synonym pairs (word-translation pairs that have at least one common synonym), illustrated in Figure 2 and in Table 2, as high as ~60%, is significant. This is encouraging for further use of this special kind of word translated pairs in tasks such as automatic enhancement of lexical databases (such as WordNet) for less resourced languages such as Romanian, based on corresponding English versions of these lexical databases.

lang A	lang B	HS % (2)
RO	FR	54,44%
FR	RO	53,57%
RO	EN	42,10%
EN	RO	46,90%
FR	EN	49,54%
EN	FR	61,94%
RO	ES	46,81%
ES	RO	41,90%
FR	ES	56,53%
ES	FR	56,83%
EN	ES	60,27%
ES	EN	52,66%

Table 2: Hard synonyms

The proportion of hard synonym pairs for each language pair can be used to gain insight into the synonym overlap of the four languages, and thus, into their degree of common concept lexicalization. The higher overlap for French-Spanish or French-Romanian (which are all latin languages) as well as for English-French (the lexicon of the English language is rich in French words) suggests that the percent of hard synonym pairs for a pair of languages could be used as a measure of lexical similarity between languages.

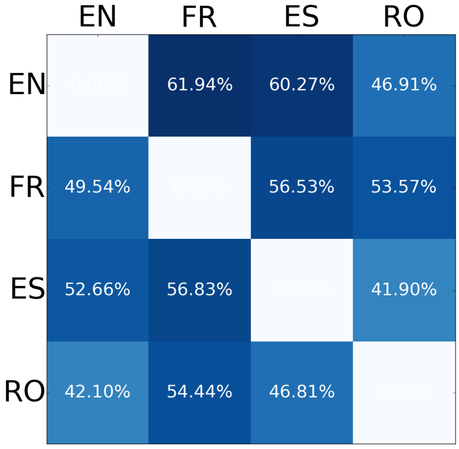


Figure 2: Hard synonym pairs percent

4. Distributional Synonymy Experiments

Distributional methods have been used successfully to automatically extract semantically similar words from corpora. Nevertheless, it is a well known problem for distributional approaches to synonym extraction that contextually similar words are not necessarily semantically similar, but sometimes merely semantically related, or even antonyms.

Among the distributional solutions for extraction of semantically similar words, multilingual approaches have been successfully used for synonym detection. (Bannard and Callison-Burch, 2005) use back-translations in parallel corpora to extract synonyms, assuming that words that translate into the same word in a pivot language are likely to be semantically similar.

We propose extending this assumption to consider words that translate into synonyms in another language. The high percentage of hard synonym pairs obtained in the experiment in the previous section, using only the dictionary synonyms and the translations provided by Google Translate, suggests it may be reasonable to assume that synonyms in one language are likely to remain synonyms in another language upon translation. This points to a potential new method for discovering new synonyms from corpora. We propose using the hard synonymy property to identify true synonyms from corpora, and distinguish between these and other distributionally similar words.

The experiment we propose consists of investigating whether distributionally similar words translated into words that are synonyms in another language, and assuming the ones that do are true synonyms rather than more weakly semantically related or antonyms.

We also investigate the effect of including more languages so as to formulate a more relaxed condition for hard synonymy: we will consider two synonyms to be hard synonyms if they maintain their synonymy upon translation into either one of two or three different languages.

4.1. Data and Methodology

We performed an experiment using as input an exhaustive list of English words from our database, obtaining a total of 44.913 input words. For each of these, we obtained distributional synonyms, by using word2vec to extract the first 100 distributionally similar words for each of the English words in our list. Using the translations and synonyms in our database, we then translated each of the distributional synonyms into a target language, and tested whether their translations are synonyms with the original word's translation in the same target language, identifying hard synonym pairs. We propose that the distributionally similar words found to be hard synonyms in the target language are likely candidates for true synonyms.

We defined a recall metric to measure how many of the hard synonyms extracted using the method above can be found as synonyms in a dictionary in the original language, using the data in our database. This measure will be used as an approximation of the likelihood that our method finds true synonyms.

If we define ds as the number of synonyms of the original word in the English dictionary, and hs as the total number of hard synonyms identified by our method, then the recall can be computed as follows:

$$recall = \frac{ds}{hs} * 100 \qquad (1)$$

4.2. Results

Using the original list of English words, and French as a target language, we obtained a recall of 40.32%, representing the percent of hard synonyms found by our method that were confirmed synonyms in the dictionary. We suggest that the rest of the extracted hard synonyms, though not in the dictionary, are still likely candidates for true synonyms.

We repeated the experiment using more than one language as a target language, by translating the distributionally similar words into two, then three languages, and testing whether their translations are synonyms with the original word's translation in any of the target languages. This significantly increases the recall up to 52,38%, suggesting our method is a reliable way to discover true synonyms among distributionally similar words.

Figure 3 below shows how the recall increases with adding more languages, illustrating the average recall obtained by using one, two and three languages respectively, for every combination of languages among French, Romanian and Spanish.

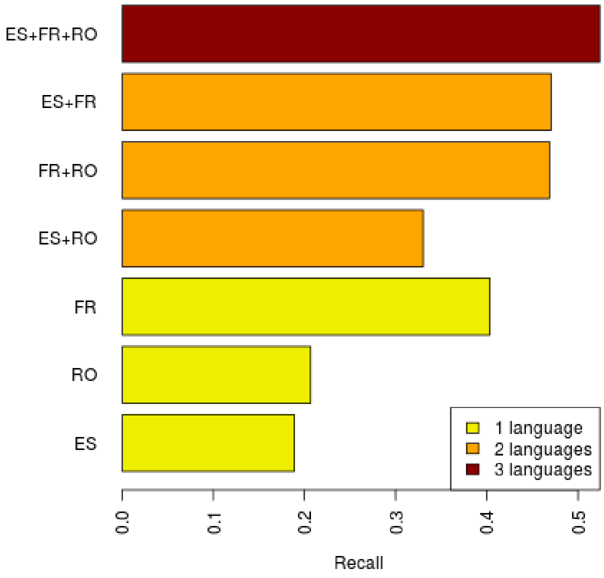


Figure 3: Synonyms recall evolution

5. Frequency of Hard Synonyms in Parallel Corpora

Hard synonym pairs (word-translation pairs that have at least one common synonym) express a common cross-lingual lexicalization of the same concept - thus we might expect a high degree of co-occurence of hard synonym word-translation pairs in a parallel corpora, in relation to non-hard synonym pairs. We conduct an experiment to test this hypothesis by measuring the relative frequencies of hard synonym pairs and non-hard synonym pairs on a parallel word-aligned corpus.

For this experiment we used the Europarl 7.0 sentence-aligned parallel corporus for English-French and English-Romanian. We used GIZA++ to align the corpus at word level for each of the language pairs, and we lemmatized all words in the corpus using DEXonline[1] for Romanian and TreeTagger for English and French. For each word-translation pair found in the word-aligned corpus, we tested whether it is a hard synonym pair: that is, whether the word and its translation have common synonyms, using our database, as described in the previous chapters. We computed the frequency of word-translation pairs that are hard synonym pairs in the aligned corpus.

The results of this experiment don't show a significant difference between the frequency of hard synonym pairs as compared to non-hard synonym pairs: the percent of hard synonym pairs is close to 50% for both language pairs, as shown in table 3.

Aligned corpus	Frequency
EN-RO	44,59%
EN-FR	52,32%

Table 3: Hard synonym pairs frequency

6. Conclusions

We have presented an analysis of the hard synonymy property and its potential applications to synonym extraction from corpora and to machine translation, performing experiments on synonyms and their translations in four languages. We have built a database containing pairs of (translated) words from the four languages along with their corresponding synonym sets and their synonym overlap set, and made it publicly available. Furthermore, we used it in order to gain insight into the synonym overlap of the four languages, and thus, into their degree of common concept lexicalization, by various queries.

We have shown that hard synonymy can be useful for improving the results of automatic synonym extraction with distributional methods, and based on these results we proposed a method for discriminating between semantically similar words (likely synonyms) and distributionally similar words that are not true synonyms. Additionally, our experiments show how increasing the number of languages considered for extracting hard synonyms increases the accuracy of the method for detecting true synonyms.

We have also investigated the potential use of hard synonym pairs for lexical machine translation, and have shown that an initial experiment on the Europarl parallel corpus doesn't support the theory that hard synonym pairs (words that have at least one common synonym with their translation) could be better candidates for lexical translation than non-hard synonym pairs.

The relative percent of synonyms overlap for each of the language pairs considered in this article suggests that it could be interesting to consider it as a measure for lexical similarity between languages. We leave for future research applying the same experiment on additional languages in order to test the validity of this theory. The relatively high percentage of hard synonym pairs (as high as ~60%) is encouraging for further use of this special kind of word translated pairs in tasks such as automatic enhancement of lexical databases (such as WordNet) for less resourced languages such as Romanian, based on the corresponding English versions.

References

Bannard, C. and Callison-Burch, C. (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 597–604. Association for Computational Linguistics.

Bikel, D. M. and Castelli, V. (2008). Event matching using the transitive closure of dependency relations. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 145–148. Association for Computational Linguistics.

Budanitsky, A. and Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

Dinu, A., Dinu, L. P., and Uban, A. S. (2015). Cross-lingual synonymy overlap. In *RANLP*, pages 147–152.

Merriam-Webster. (2004). *Merriam-Webster's collegiate dictionary*. Merriam-Webster.

Mohammad, S. and Hirst, G. (2006). Distributional measures of concept-distance: A task-oriented evaluation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 35–43. Association for Computational Linguistics.

Mohammad, S., Gurevych, I., Hirst, G., and Zesch, T. (2007). Cross-lingual distributional profiles of concepts for measuring semantic distance. In *EMNLP-CoNLL*, pages 571–580.

Van der Plas, L., Merlo, P., and Henderson, J. (2011). Scaling up automatic cross-lingual semantic role annotation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 299–304. Association for Computational Linguistics.

Wang, T. and Hirst, G. (2011). Refining the notions of depth and density in wordnet-based semantic similarity measures. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1003–1011. Association for Computational Linguistics.

Wang, W., Thomas, C., Sheth, A., and Chan, V. (2010). Pattern-based synonym and antonym extraction. In *Proceedings of the 48th Annual Southeast Regional Conference*, page 64. ACM.

[1] http://dexonline.ro

Towards Preparation of the Second BUCC Shared Task: Detecting Parallel Sentences in Comparable Corpora

Pierre Zweigenbaum[1] **Serge Sharoff**[2] **Reinhard Rapp**[3]
[1]LIMSI, CNRS, Université Paris-Saclay, Orsay, France
[2]University of Leeds, United Kingdom
[3]University of Mainz, Germany
pz@limsi.fr s.sharoff@leeds.ac.uk reinhardrapp@gmx.de

Abstract

In this paper we provide a summary of the rationale and the dataset contributing to the second shared task of the BUCC workshop. The shared task is aimed at detecting the best candidates for parallel sentences in a large text collection. The dataset for the shared task is based on a careful mix of parallel and non-parallel corpora. It contains 1.4 million French sentences and 1.9 million English sentences, in which 17 thousand sentence pairs are known to be parallel. The shared task itself is scheduled for the 2017 edition of the workshop.

Keywords: Parallel corpora, cross-language similarity

1. Introduction

Shared tasks gained importance in the NLP community since the data turn in the 1990s. They provide a way to compare different approaches using a common dataset and evaluation methods. In the field of comparable corpora, there are several options for shared tasks, such as:

1. methods for collecting comparable corpora from the Web;
2. methods for assessing the similarity of documents across languages in a collection of texts;
3. methods for assessing the similarity of separate sentences across languages in comparable corpora;
4. methods for detecting translations of words and phrases across languages in comparable corpora.

While it is difficult to operationalise the first task in this list, the 2015 edition of the BUCC workshop included the second task from this list (Sharoff et al., 2015). In it we used aligned Wikipedia articles to test document-level comparability methods for linking Chinese, French, German, Russian and Turkish articles to English.

In 2016 we aimed at building resources to test sentence-level comparability approaches. This paper describes our rationale for designing these resources, the methods used to build them, and the resulting data. A shared task based on these resources is planned for BUCC 2017.

2. Objectives

Our objectives were to create a dataset to evaluate parallel sentence extraction from comparable corpora.

Most former research on parallel sentence extraction from comparable corpora has relied on specific properties of the corpora used. This includes date properties in synchronous comparable corpora, e.g., international news in the same range of dates (Utiyama and Isahara, 2003; Munteanu et al., 2004; Abdul-Rauf and Schwenk, 2009), or document-level parallelism, e.g., encyclopedia articles for matched entries in two languages, as in Wikipedia.

The dependency on these specific properties creates two problems in our opinion. At a first level, we observe that these properties vary with the addressed corpora, and that they add to the difficulty of assessing the behavior of parallel sentence extraction methods. At a deeper level, we consider that the 'pure' task of translation spotting in comparable corpora should focus on content-based properties of the texts, not on external metadata.

The objective of the targeted task is therefore to test the ability of methods to detect parallel sentences in pairs of monolingual corpora *without using any metadata* on the corpora. In this task, only intrinsic properties of the sentences can be used.

Our initial design includes the following criteria, which we further refine and complement below after a study of related work:

- We start from two comparable corpora: these should not be the result of translations, as far as possible.
- No structural clues are provided beyond the order of sentences, which aims to be natural: the dataset provides no pre-existing document alignment (as in date-synchronized news or in linked Wikipedia pages).
- To be able to evaluate systems which detect parallel sentences in this pair of corpora, we need to know all 'positive examples' of parallel sentence pairs in these corpora. Therefore, we decided to introduce known pairs of parallel sentences into these comparable corpora.

3. Related work

This section briefly reviews related work relevant to the preparation of a corpus for the detection of parallel sentences.

3.1. Plagiarism detection: PAN

Shared tasks on plagiarism detection, as embodied by the PAN series (e.g., Potthast et al. (2012)), aim to detect instances of 'text re-use': text borrowed from one text into another. From the first editions on, PAN datasets have included not only monolingual but also cross-language instances of text re-use (Potthast et al., 2011).

The problem of detecting cross-language text re-use can be formulated as follows: does a text re-use parts of a previous text in a different language? It can be addressed as an 'intrinsic' cross-language plagiarism detection task, where 'translationese' is differentiated from original language (Barrón Cedeño, 2012, p. 145): methods for monolingual plagiarism detection can apply, such as differences in the distribution of function words or in language models. What Barrón Cedeño (2012, p. 147) names 'external' cross-language plagiarism detection is equivalent to the task of detecting text fragments with a high level of comparability (in particular parallel and highly comparable) from a multilingual corpus. In other words, we could consider that external cross-language text re-use and text alignment can be addressed as the same task, viewed from two different perspectives. Barrón Cedeño (2012) outlines five types of methods:

1. Models based on 'Syntax' (actually, morphology):
   Character dot-plot
   Character n-grams
   Cognateness
2. Models based on Thesauri (= single-word or term translation):
   EuroWordNet thesaurus
   Eurovoc thesaurus
3. Models based on Comparable Corpora (actually, aligned non-translated documents, namely Wikipedia)
   Cross-language explicit semantic analysis
4. Models based on Parallel Corpora:
   Bilingual representation space: Cross-language latent semantic indexing
   Bilingual mapping: Cross-language kernel canonical correlation analysis
5. Models based on Machine Translation (MT): Language normalisation (i.e., translation into one language)
   Web-based cross-language models (same as above, using on-line MT service)
   Multiple translations (i.e., output of MT before language model, with multiple translation hypotheses)

The present BUCC task is different from the PAN cross-language plagiarism detection in the following ways:

- The BUCC task aims to evaluate 'external' cross-language detection, whereas PAN is interested in both 'intrinsic' and 'external' cross-language plagiarism detection. As a consequence, the BUCC dataset should reduce the ease with which intrinsic plagiarism detection methods could spot artificially introduced sentences.

- The BUCC task focuses on sentence-level text fragments, whereas this granularity is not required by PAN.

3.2. Semantic text similarity: SemEval 2016

Semantic text similarity assesses the semantic equivalence of two texts or text fragments, e.g. sentences. Cross-language sentence similarity is close to evaluating whether two sentences are translations of one another: if they are, they obtain maximum similarity.

SemEval 2016[1] includes a cross-language sentence similarity task: its goal is to evaluate the similarity of sentence pairs which belong to two different languages, instantiated on the English-Spanish language pair. The task is formulated as scoring a given pair of sentences on a six-point scale.

The trial data was drawn from sentence pairs used in prior English semantic text similarity evaluations (STS 2012, 2013, 2014, and 2015). Bilingual data was obtained by translating some of the English sentences into Spanish and considering that the semantic similarity score for a resulting cross-lingual pair was that of the original English sentence pair, then filtering out some lower quality cross-lingual pairs.

Note that the interpretation of the scores can be related to comparability. The examples provided by the organizers and their explanations of the scores are copied below, from highest to lowest similarity *(We added an English translation of the Spanish sentences in parentheses.)*. The highest similarity score (5) corresponds to an exact translation, the next (4) is probably an acceptable translation, whereas the following one (3) would be an inexact translation and would be likely to obtain a lower BLEU score. Sentence pairs with lower scores would be likely to introduce too much noise if added to the training corpus of a statistical machine translation system.

(5) The two sentences are completely equivalent, as they mean the same thing
El pájaro se está bañando en el lavabo. *(The bird is washing itself in the water basin.)*
Birdie is washing itself in the water basin.

(4) The two sentences are mostly equivalent, but some unimportant details differ.
En mayo de 2010, las tropas intentaron invadir Kabul. *(In May 2010, the troops tried to invade Kabul.)*
The US army invaded Kabul on May 7th last year, 2010.

(3) The two sentences are roughly equivalent, but some important information differs or is missing.
John dijo que él es considerado como testigo, y no como sospechoso. *(John said that he is considered as a witness, not as a suspect..)*
"He is not a suspect anymore." John said.

(2) The two sentences are not equivalent, but share some details.
Ellos volaron del nido en grupos. *(They flew from the nest in groups.)*
They flew into the nest together.

(1) The two sentences are not equivalent, but are on the same topic.
La mujer está tocando el violín. *(The woman is playing the violin.)*
The young lady enjoys listening to the guitar.

(0) The two sentences are on different topics.
Al amanecer, Juan se fue a montar a caballo con un grupo de amigos. *(At dawn, Juan went riding with a group of friends.)*
Sunrise at dawn is a magnificent view to take in if you wake up early enough for it.

[1] http://alt.qcri.org/semeval2016/task1/

The README of the task suggests methods to compute cross-language similarity: Adapting monolingual 'align+featurize' semantic text similarity systems to the cross-lingual task; deep learning with cross-lingual embeddings; and monolingual semantic text similarity complemented with machine translation.

The present BUCC task has two differences with cross-language semantic text similarity:

- The BUCC task uses a binary scale to evaluate whether or not two sentences are translations of each other.
- The BUCC task does not provide a list of sentence pairs, but instead provides two monolingual lists of sentences. The set of sentence pairs to be examined by the systems is potentially the cross-product of these two sets of sentences: this creates the need for efficient comparison or pruning methods.

3.3. Bilingual document alignment: WMT 2016

WMT 2016 includes a shared task on bilingual document alignment[2]. In that task, given two sets of Web pages in two languages from the same Web domain, each pair of translated source-target page pairs must be detected.

Whereas the similarity to the BUCC task is clear, two differences can be noted:

- The main difference is the granularity of the documents to be aligned: the BUCC task addresses sentences, whereas WMT 2016 addresses documents (Web pages).
- Another difference however is that the BUCC task aims not to use any metadata; in contrast, WMT provides metadata on its documents: the Web page URLs, which make it possible to use non-content-based methods to address the task. As a matter of fact, an implementation of such a method is provided as a baseline by the organizers and can be downloaded from the WMT Web site.

4. Data preparation methods

An ecologically sound way to produce resources for our task would be by annotating manually parallel sentences in a large selection of sentences from a real comparable corpus, i.e., two comparable monolingual corpora. However, exact translations are very rare in a randomly collected corpus, and manually spotting them would be labor-intensive. Often they imply that their two provenant documents are reasonable translations (in either direction). To increase the probability of finding parallel sentences, the two corpora could thus be selected so that they consist of pairs of matching documents on the same topic. But many sentences in a collection of aligned Web documents are likely to originate from machine translated texts (Antonova and Misyurev, 2011). Additionally, detecting automatically translated sentences is easy if using the same MT system (primarily, Google Translate) (Potthast et al., 2012). Therefore, we switched to creating a dataset which is prepared automatically from a known parallel corpus and known non-parallel sentences from two monolingual corpora. In this dataset, known pairs of parallel sentences are 'planted' into existing monolingual corpora.

The above-mentioned work on cross-lingual plagiarism suggests to invest some effort into a reasonably good blend of the planted sentences in their environment (what the plagiarism literature calls 'obfuscation'). Otherwise, it could be easier to identify which (parallel) sentences were added to the initial texts than to check their parallelism. To determine in which document a passage of another document can be inserted, Asghari et al. (2015) perform sentence and document clustering based on the sentence similarity obtained through Information Retrieval queries with the Lucene IR engine. We followed a similar though simpler approach to determine where to insert parallel sentences, which we describe below.

1. Indexing collections of monolingual sentences.

- Our initial data is composed of a pair of comparable monolingual corpora (Wikipedia dumps in two languages, say EN and FR) and a sentence-aligned parallel corpus in the same pairs of languages (News Commentary[3]).
- We split each of the two monolingual corpora into sentences (using the Europarl sentence splitter).
- We treated each monolingual corpus as a collection of sentences and indexed them with an information retrieval engine (Apache SolR[4]).

2. Spotting similar sentences through IR queries.

- For each pair of parallel sentences, we used the EN sentence as a query to the EN collection of sentences and the FR sentence as a query to the FR collection of sentences. If successful, this should identify a locations in the EN (resp. FR) monolingual corpus where the EN (resp. FR) parallel sentence could be inserted in a context where they have chances to be related to the current topic.
- Our motivation for using an IR engine is to implement a scalable sentence similarity computation process with minimal investment. We chose query parameters which impose stronger similarity constraints on the query (parallel sentence) and the 'document' (monolingual sentence), for instance by setting a minimum number of common content words (5) between query and document and imposing a similar total number of content words in both sentences.
- We consider a pair of queries as successful if it retrieves at least one EN sentence and one FR sentence with the chosen constraints.

3. Inserting parallel and non-parallel sentences into the monolingual corpora.

[2] http://www.statmt.org/wmt16/bilingual-task.html

[3] http://www.casmacat.eu/corpus/news-commentary.html

[4] http://lucene.apache.org/solr/

- Given a pair of locations (similar monolingual sentences) identified by a successful pair of queries built from a pair of aligned EN-FR sentences, we insert the parallel EN sentence after the monolingual EN sentence similar to it, and the parallel FR sentence before the monolingual FR sentence similar to it.

- After the previous step, each sentence inserted into one of the two monolingual corpora is parallel to a sentence inserted in the other monolingual corpus. This means that if a system detects inserted sentences (for instance through intrinsic plagiarism detection methods as mentioned in Section 3.1.), it can be certain that this sentence belongs to the gold standard. To reduce this certainty, we also insert A adjacent sentences from the parallel corpora together with the parallel EN and FR sentences: A sentences following the EN sentence and A sentences preceding the FR sentence, so that these added sentences are not parallel. Now not all inserted sentences are parallel anymore.[5]

4. Increasing the rate of parallel sentences.

   - In the above-described process, only a small proportion of sentences in the original comparable corpora become an insertion point for parallel sentences. To increase the rate of parallel sentences in the resulting corpus, we only keep monolingual documents (Wikipedia pages) where at least one insertion point has been found.

   - When a monolingual document is included in the corpus, if there is an interlanguage link from it to a document (Wikipedia page) in the other language, it is inserted too, even though no parallel sentence may have been inserted into that linked document.

   - Some monolingual documents are much longer than the others: to reduce the non-parallel part of the corpus further, we truncate them to their first 500 sentences.

5. Reducing the rate of non-inserted parallel sentences.

   - There is always a chance that naturally-occurring parallel sentences exist in a pair of Wikipedia pages. We need to know about them to be able to provide a fair evaluation of translation spotting systems. However detecting them automatically is the very goal of our target shared task, so we cannot assume we have a system which will do this perfectly. We envision several methods to reduce these pairs of naturally-occurring parallel sentences.

     1. Use an existing system to spot them and either add them to the gold standard or remove them from the data. A problem is that this will bias the corpus towards this system.
     2. Use an existing method or system with relaxed constraints to increase the recall of the detection of potentially parallel sentences, for instance by translating source sentences automatically to the target language and using a semantic text similarity metric (see Section 3.2.) to spot (and remove) pairs of sentences with a similarity above some relatively low similarity threshold (e.g., between 2 and 3 on the SemEval scale presented in Section 3.2.). A problem is that this will bias the distribution of cross-lingual sentence similarity, creating a gap between unrelated sentences and (inserted) translated sentences.
     3. At evaluation time, pool the results of the participating systems and have humans examine false-positive sentence pairs found by a consensus of at least N systems. This requires a human investment which remains to be estimated.

   - Since pairs of shorter sentences are more likely to be chance translations of each other, we removed from the corpus sentences with less than a ceiling of C content words.

5. Dataset

We instantiated the above-mentioned method on the French-English pair of languages:

- The monolingual corpora are July-August 2014 XML Wikipedia dumps provided by the LinguaTools Web site[6]. We prepared the text versions of these corpora by using the associated tool xml2txt[7]. HTML entities were converted into their UTF-8 equivalent. Documents were further tokenized[8] and split into sentences as detailed above. The English corpus contains 4.5M articles and 138M sentences, the French corpus 1.5M articles and 46M sentences.

- The parallel corpus comes from the News Commentary, version 9, provided as training data for WMT 2014[9]. The French-English News Commentary corpus contains 183k sentence pairs.

- After some experiments, we set the following Solr query parameters: efType=”edismax”, qs=5, ps=5, ps2=5, mm=”70%”, stopwords=”true”. With these parameters, the process retrieved similar sentences for 18k sentence pairs, representing 10% of the News Commentary sentence pairs and 0.03% of the French Wikipedia sentences.

- After completion of the process, the produced comparable corpora contain respectively 1.4M French sentences and 1.9M English sentences, including 17k inserted parallel sentences in each corpus.

[5] Indeed a system using intrinsic plagiarism detection methods might probably still spot the inserted passages and reduce the complexity of the search for parallel sentences. Again, this is not what the present shared task aims to evaluate.

[6] http://linguatools.org/tools/corpora/wikipedia-monolingual-corpora/

[7] xml2txt.pl -articles -nomath -notables -nodisambig

[8] Tokenization is performed by the Solr indexer anyway and was not really necessary at this step.

[9] http://www.statmt.org/wmt14/training-parallel-nc-v9.tgz

French monolingual sentences	English monolingual sentences
fr-000000197 Si bien que l'année suivante, elle mit sa priorité dans les initiatives régionales telles que le Mercosur ou la Banque du Sud après une décennie de partenariat avec les États-Unis. **fr-000000198** Prenons l'exemple du MERCOSUR (le Marché commun du Sud), la principale initiative régionale d'après-guerre. **fr-000000199** Selon l'universitaire argentin Roberto Bouzaq, le MERCOSUR est dans un état critique en raison de son incapacité à maintenir le cap sur les objectifs communs qui ont conduit les pays-membres à s'engager dans un processus d'intégration régionale, avec pour conséquence un éparpillement et l'impossibilité d'identifier les problèmes politiques sous-jacents qui devraient être prioritaires. ... **fr-000000203** Enfin, l'Argentine fut l'un des signataires initiaux du Traité sur l'Antarctique. **fr-000000204** Enfin, l'Argentine est un cas à part.	**en-001425664** All the while, scant attention is paid to the region's already established bodies, which are in sad shape. **en-001425665** Consider MERCOSUR, the main post-Cold War regional initiative. ... **en-001876436** Indeed, this vision of international relations clearly rests on building influence through military power. **en-001876437** Finally, Argentina stands in a category by itself.

Table 1: Example sentences: comparable corpora with inserted pairs of parallel sentences (see Table 2).

Table 1 shows an excerpt of our collection. Two out of four sentences in each language are linked in the gold alignment file, as shown in Table 2.

fr-000000198	⇔	en-001425665
fr-000000204	⇔	en-001876437

Table 2: Example gold standard alignments (sentence pairs from parallel corpus).

6. Limitations

The current design and its realization have the following limitations.

- The insertion of the parallel sentence pairs (from News Commentary) into the monolingual corpora (from Wikipedia) is sometimes coherent, sometimes not really coherent.
- At some point in the implementation of the method the monolingual corpora were tokenized, but not the parallel corpora. This created surface differences which can reveal the origin of sentences. 'Detokenizing' (pasting back punctuation to the adjacent token) is not an easy process, and we should reprocess the corpus without tokenization, which was not really needed in our pipeline.
- The need for obfuscating the inserted parallel sentence pairs remains a matter of debate. A much higher quality would be required of the blending of inserted sentences into the monolingual corpora than what was performed here, for instance as in (Asghari et al., 2015), for it to be really useful.
- Translation pairs that may exist naturally in Wikipedia are not removed nor known exactly, and may hence lead to counting false positives in the evaluation if systems find them. Human review of pooled system results are a possible solution to this problem, but require manpower.
- The above-described method was applied to the French-English language pair as a proof of concept. It is yet to be applied to other language pairs. This would be feasible in principle for German, Russian, and Chinese, for which source data are available in both Wikipedia and News Commentary (and which Solr can handle). Turkish is handled by Solr but is not present in the News Commentary collection of parallel corpora.

7. Evaluation method

The primary evaluation measure is the F-score of sentence pairs:

- A sentence pair is considered correct if it is present in the gold standard.
- Precision is the proportion of correct system-generated pairs among those pairs returned by the system.
- Recall is the proportion of correct system-generated pairs among all pairs in the gold standard.
- F is the harmonic mean of precision and recall.

8. Shared task plans

Because of the complexities involved in preparation of the dataset, the task initially proposed for the 2016 edition of the BUCC workshop had to be postponed to 2017.

References

Abdul-Rauf, S. and Schwenk, H. (2009). Exploiting comparable corpora with TER and TERp. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: From Parallel to Non-parallel Corpora*, BUCC '09, pages 46–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Antonova, A. and Misyurev, A. (2011). Building a web-based parallel corpus and filtering out machine-translated text. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 136–144.

Asghari, H., Khoshnava, K., Fatemi, O., and Faili, H. (2015). Developing bilingual plagiarism detection corpus using sentence aligned parallel corpus. In *Notebook for PAN at CLEF 2015*, Toulouse, France.

Barrón Cedeño, L. A. (2012). *On the Mono- and Cross-Language Detection of Text Re-Use and Plagiarism*. Ph.D. thesis, Universitat Politècnica de València.

Munteanu, D. S., Fraser, A., and Marcu, D. (2004). Improved machine translation performance via parallel sentence extraction from comparable corpora. In Daniel Marcu Susan Dumais et al., editors, *HLT-NAACL 2004: Main Proceedings*, pages 265–272, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

Potthast, M., Barrón-Cedeño, A., Stein, B., and Rosso, P. (2011). Cross-language plagiarism detection. *Language Resources and Evaluation*, 45(1):45–62.

Potthast, M., Gollub, T., Hagen, M., Kiesel, J., Michel, M., Oberländer, A., Tippmann, M., Barrón-Cedeño, A., Gupta, P., Rosso, P., and Stein, B. (2012). Overview of the 4th international competition on plagiarism detection. In Pamela Forner, et al., editors, *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*, volume 1178 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Sharoff, S., Zweigenbaum, P., and Rapp, R. (2015). BUCC shared task: Cross-language document similarity. In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, pages 74–78, Beijing, China, July. Association for Computational Linguistics.

Utiyama, M. and Isahara, H. (2003). Reliable measures for aligning japanese-english news articles and sentences. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 72–79, Sapporo, Japan, July. Association for Computational Linguistics.