Identifying Cross Language Term Equivalents Using Statistical Machine Translation and Distributional Association Measures

Hans Hjelm
CL Group, Department of Linguistics
Graduate School of Language Technology (GSLT) and Stockholm University
SE-106 91 Stockholm, Sweden
hans.hjelm@ling.su.se

Abstract

This article presents a comparison of the accuracy of a number of different approaches for identifying cross language term equivalents (translations). The methods investigated are on the one hand associative measures, commonly used in word-space models or in Information Retrieval and on the other hand a Statistical Machine Translation (SMT) approach. I have performed tests on six language pairs, using the JRC-Acquis parallel corpus as training material and Eurovoc as a gold standard. The SMT approach is shown to be more effective than the associative measures. The best results are achieved by taking a weighted average of the scores of the SMT approach and disparate associative measures.

1 Introduction

This article deals with the identification of cross language term equivalents, a topic interesting for its applicability in a number of language technology fields. The most obvious application is the automatic construction of domain-specific bilingual dictionaries. Such dictionaries are used in many different settings, including e.g., rule-based Machine Translation and Computer-Assisted Language Learning. Some approaches in Cross Language Information Retrieval also rely on the existence of bilingual dictionaries, for translating queries. The research presented in this article also integrates into Ontology Learning; this is described in section 3.

Many researchers have proposed various kinds of distributional association methods for the bilingual dictionary extraction task, see e.g., (Church and Gale, 1991), (Fung and Church, 1994) and (Smadja et al., 1996). Other researchers have tried to solve the task by using methods from SMT, see e.g., (Melamed, 2000) and (Tsuji and Kageura, 2004), though the focus there is word alignment rather than dictionary extraction.

This article presents a systematic comparison of these two main approaches on a variety of language pairs, using the JRC-Acquis parallel corpus (Steinberger et al., 2006) to train the models, and Eurovoc V4.2[1] to evaluate the results. Contrary to what is reported in (Sahlgren and Karlgren, 2005), the SMT approach here outperforms the associative measures. I also show that the results from the SMT approach can be improved by weighting them together with the results from the associative measures in an ensemble approach.

2 Background

This section gives a brief overview of related work using the associative measures and the SMT approach separately, followed by attempts to combine the two.

2.1 Distributional association measures

A number of articles have been published during the past two decades, where the distributional characteristics of words or terms in natural language texts

[1] http://eurovoc.europa.eu/

Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek and Mare Koit (Eds.)
NODALIDA 2007 Conference Proceedings, pp. 97–104

have been exploited in order to measure the semantic similarity between those same words or terms. In (Sahlgren, 2006), a major distinction is drawn between *syntagmatic* and *paradigmatic* relations. Words that stand in a syntagmatic relation to each other are words like *cradle – baby*; there is a thematic connection, but the two words do not necessarily share many semantic features. Conversely, the words *cradle – bed* are paradigmatically related, and many more semantic features are shared. Lund and Burgess (1996) refer to these relations as *associative* and *semantic*, respectively. Sahlgren also links syntagmatic relations to information contained in term-document co-occurrence models and paradigmatic relations to term-term co-occurrence models.

When dealing with large amounts of text, on the order of giga- or terabyte, calculations on co-occurrence matrices become very expensive, regarding both resources and time. To bypass this problem, different methods for reducing the dimensionality of the matrices have been proposed. In (Sahlgren and Karlgren, 2005), Random Indexing[2] is used for this very purpose, and the result is evaluated on a bilingual lexical acquisition task. Another widely used method for dimensionality reduction, the singular value decomposition (see e.g. (Golub and van Loan, 1996)), has yet to be evaluated on the dictionary extraction task. I hope to report on the results of ongoing experiments in this direction in the near future.

Regardless of whether dimensionality reduction has been performed or not, each word or term (row in the matrix) can be compared to each other word or term, using similarity measures defined for vectors. There is a plethora of such measures, many of which have been evaluated on the present task, or one similar to it. In (Ribeiro et al., 2000), a total of 28 different similarity measures are evaluated on extracting equivalents from aligned parallel texts. The task is similar to the one presented here, but they use one language pair (Spanish – Portuguese) for testing on a parallel corpus containing about 18,000 words. Two of the highest ranking measures in that evaluation, the cosine measure and the Mutual Information measure,[3] are compared in section 3.

How to exploit distributional models for solving the task at hand is described more closely in sections 3.3.1 through 3.3.3.

2.2 Statistical Machine Translation (SMT)

GIZA++,[4] which builds on IBM's translation models 1–5 (Brown et al., 1993), produces a bilingual dictionary file, where each source language word or term is listed with its possible translations and associated probabilities. The most probable translation of a particular source term can thus be found by sorting the possible translations in descending order based on their associated probabilities and then selecting the first translation in the sorted list. Melamed (2000) describes three statistically based approaches, all making use of co-occurrence information coupled with e.g., a noise model or statistical smoothing. Och and Ney (2003) propose extensions of IBM's translation models and show improvements on a word alignment task; the system is not evaluated on a dictionary extraction task.

2.3 Combining distributional association measures with SMT

Tiedemann (2003) proposes a method for word alignment which makes use of both distributional association measures[5] and the dictionary files produced by GIZA++ mentioned above. Note that the evaluations performed there are on a token level, rather than on a type level, which is what we are interested in here. Tiedemann also uses other information, such as string matching and part of speech, and so is able to boost the performance of GIZA++ by weighting the scores of the different sources together. However, in at least one of Tiedemann's evaluations, including information from any other source than GIZA++ resulted in a decrease in system performance.

3 Experimental setup and results

I compare the results for the distributional models when varying three different parameters:

1. Whether a matrix containing co-occurrence information based on shared neighbors (paradig-

[2] See the quoted article for a description of the Random Indexing methodology.

[3] Referred to as *Average Mutual Information* in Ribeiro's evaluation.

[4] http://www-i6.informatik.rwth-aachen.de/web/Software/

[5] He refers to these measures as *co-occurrence measures*.

matic) or shared documents/text segments (syntagmatic) is used.

2. Whether Random Indexing or no dimensionality reduction is used.

3. Whether cosine or Mutual Information is used as the similarity measure.

I describe each alternative further in sections 3.3.1 through 3.3.3.

3.1 Translating terms

Why translate terms rather than words? Consider e.g., ontologies and Ontology Learning (see e.g., (Cimiano, 2006)), a field growing in importance along with the emergence of the Semantic Web. In Ontology Learning, one of the main tasks is to identify all expressions that are of particular importance within the domain of interest, e.g., medicine or law. These expressions can consist of a single word or they can be multi-word units. When we are looking at a particular domain, these expressions are assumed to correspond to the terms in that domain. A lot of work in the field of Term Extraction has been carried out towards automating the term extraction process (see e.g., (Castellví et al., 2001; Jacquemin, 2001)).

After term extraction, the next question of interest for an ontology engineer would be whether some of the extracted terms refer to the same *concept*. A concept, as I use the term here, is compatible with the topmost point in Peirce's semiotic triangle (Ogden and Richards, 1923), connecting symbols (here terms) with objects or phenomena in the real world. Roughly, if we are dealing with terms from the same language that refer to the same concept, we say that these terms are *synonyms*. If the terms are from different languages, we call them *equivalents*. It is the latter that I am interested in identifying in this study. In the ontology learning application scenario, we are interested in finding equivalence relations between *terms* in the source and target languages – relations between terms and non-terms ("regular words") or relations purely between non-terms are only of secondary interest.

In my experiments, I assume that the term extraction has already been carried out correctly. This means two things:

1. The task for the systems consists in translating the Eurovoc terms.

2. The translation candidates are limited to the target language terms – no non-terms are allowed as translation candidates.

This may seem like a rigid restriction. However, if we assume that the term extraction process has been carried out correctly and we also assume that a *term* in the source language is always translated with a *term* in the target language, this restriction is needed for sake of consistency.

3.2 Data and gold standard

I used the JRC-Acquis parallel corpus for building the distributional models and for training the GIZA++ system. The corpus consists of legal texts concerning matters involving the EU. I have used all pairwise combinations of the following languages in my experiments: German, English, French and Swedish. This means that six language pairs have been evaluated and thus twelve directions of translation. The number of words per language varies between 6.5 million (Swedish) and 7.8 million (French).

The parallel corpora are distributed in a format where they have been aligned automatically on a paragraph level. The paragraphs are very short and usually only contain one sentence or even one part of a sentence. There are two alignment versions available for download;[6] I used the version produced by the Vanilla aligner[7] in my experiments. To ease some of the usual problems caused by sparse data (which is even worse when working with terms than with words), I lemmatized the texts using Intrafind's[8] LiSa system for morphological analysis (Hjelm and Schwarz, 2006). The Swedish texts, though, had to be left unprocessed, due to a lack of resources.

As a gold standard, against which to check the translations proposed by the system, I used Eurovoc V4.2, a freely available multilingual thesaurus existing in more than 20 languages and covering topics where the EU is active. The thesaurus con-

[6] http://wt.jrc.it/lt/Acquis/
[7] http://nl.ijs.si/telri/Vanilla/doc/ljubljana/
[8] http://www.intrafind.de

tains 6,645 concepts, each of which is given a *descriptor*, or recommended term, in each language. These descriptors constitute my gold standard; when the system translates the descriptor for a concept in the source language with the descriptor for the same concept in the target language, the translation is counted as correct, otherwise as incorrect. I also lemmatized the descriptors, in order for the gold standard to be on the same format as the corpora.

Next, I applied a very simple term spotting technique (for more on term spotting, see (Jacquemin, 2001)). Going through each text from left to right, I simply marked the longest matching string of complete words, that also is a descriptor for the language in question, as a term. I marked the terms so that they would be recognizable and so that the system would be able to treat them as single textual units. For example:

```
A new accounting system was installed. ⇒
a new ACTERM_accounting_system#4362 be
install .
```

3.3 Comparing the distributional models

Throughout all experiments, I use the log_2 of the frequencies in the models rather than using raw frequencies. The intuition behind this is that a word co-occurring twice with another word should be weighted higher than a word that co-occurs only once – but probably not *twice* as high. In Information Retrieval, using log frequencies, or the *logarithmic term frequency*, is a standard technique. It has also been applied successfully e.g., to the closely related problem of automatic thesaurus discovery (Grefenstette, 1994).

The matrix rows are then normalized so that the vectors are of unit length, in preparation for using the cosine measure.

3.3.1 Syntagmatic vs. paradigmatic models

When building the syntagmatic model, rows represent terms and columns represent documents, or in this case paragraphs. One model per language and language pair is needed, since the paragraph alignment is unique to each language pair.

When building a paradigmatic model, one usually makes use of a fixed-size sliding window to determine which words are to be considered neighbors of the focus word. In these experiments, I use the target language part of the alignment unit as the window, as illustrated in figure 1. Nothing actually forces us to use the *target* language words as features, we might as well use the *source* language words as features, or use both. I will return to this point in section 3.3.4. I make no adjustment for the proximity of the words, since I do not wish to make any assumptions about the similarity of word order between the languages involved.

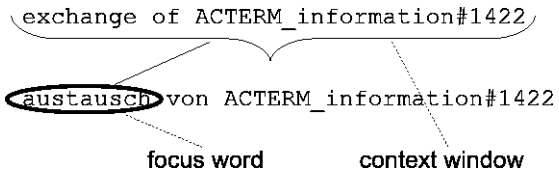

Figure 1: Constructing the paradigmatic model for translating from German to English. The focus word is circled.

3.3.2 Random indexing vs. full matrix

As mentioned previously, I wanted to compare the effects of using no dimensionality reduction with that of using Random Indexing. Of course, using a reduced matrix can give computational benefits (see section 2.1), especially when working with larger text collections. Here, I am mainly interested in the effects it might have on the *accuracy* of the system.

3.3.3 Cosine vs. Mutual Information

It would have been methodologically pleasing to try the different kinds of similarity measures with all combinations of syntagmatic vs. paradigmatic models, paired with both options for dimensionality reduction named previously. However, applying the Mutual Information measure does not make sense after the dimensionality reduction has been performed, since most or all vectors will be dense by then, containing few or no zeros. I use the following following formula to calculate Mutual Information:

$$\sum_{x,y} p(x,y) log \frac{p(x,y)}{p(x)p(y)}$$

This typically presupposes a binary representation, meaning that if the number of zero-entries in all vectors is very low or zero, the measure will judge most

or all vectors to be equally similar to each other. I therefore refrained from evaluating the Mutual Information measure on models where dimensionality reduction had been performed. For the cosine measure, since the vectors are of unit length, I only have to calculate the dot product between the two vectors.

Given the great number of similarity measures available, it would have been possible to include many more in the evaluation. The cosine measure was chosen because of its widespread application in Information Retrieval and the Mutual Information measure because of its acceptance in the Information Theory community along with its giving the best results in the comparison in (Ribeiro et al., 2000).

3.3.4 Results for the comparison of the distributional models

For each combination of settings, I evaluated each of the twelve translation directions. Again, as mentioned in section 3, I only consider the descriptors of the target language as translation candidates. As input to the system, I use all source language descriptors that occur at least once in the source language text of the parallel corpus at hand. I also split the descriptors into eleven frequency classes (counted separately for each of the twelve directions of translation): 1, 2–5, 6–10, 11–50, 51–100, 101–500, 501–1000, 1001–5000, 5001–10000, 10001–50000 and 50001<=. I calculated the average accuracy for all twelve directions of translation, for each frequency class as well as the overall accuracy, regardless of frequency (displayed later in table 1). Figure 2 shows a comparison of all applicable combinations of settings when working with paradigmatic models. Figure 3 shows the same comparison for the syntagmatic models.

As mentioned in section 3.3.1, there is no inherent reason to choose the target language words as features when building a paradigmatic model. In fact, since four languages were involved in these experiments, I made an experiment where words from all four languages were used as features. As can be seen in table 1 (where this method is labeled "Paradigm-Full-Cosine-CL"), this brought a very moderate increase in performance, but still makes this the most effective paradigmatic model.

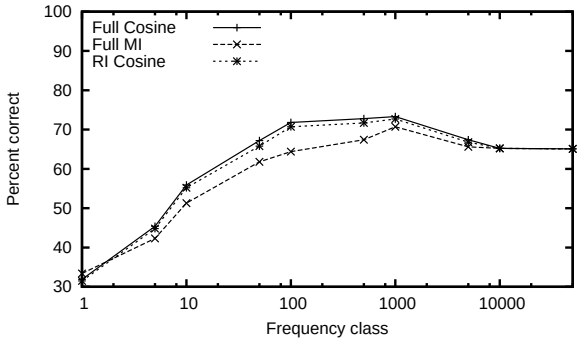

Figure 2: Paradigmatic models. "Full" stands for no dimensionality reduction, "MI" for Mutual Information and "RI" for Random Indexing

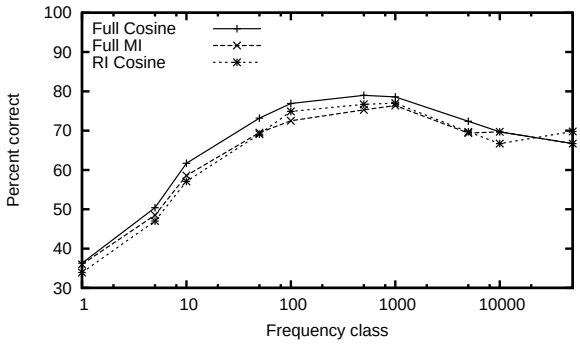

Figure 3: Syntagmatic models.

3.4 SMT vs. the distributional models

I ran the GIZA++ system with the standard settings provided in the publicly available distribution. Since all terms are treated as single words by the system, after the term spotting applied during preprocessing, we sidestep the problem of the lacking possibility in GIZA++ of capturing many-to-many relations. Figure 4 displays a comparison between the best performing syntagmatic and paradigmatic models with the results from GIZA++. "CL" in the figure stands for "Cross Language" and refers to the fact that words from all four languages involved were used as features when training that model.

3.5 Ensemble method

I combined the results of the top performing models, shown in figure 4, in an ensemble method. The idea here is that, even though the statistical model outperforms the other two, they may still contain useful information that the statistical model is missing. There are at least two factors one would

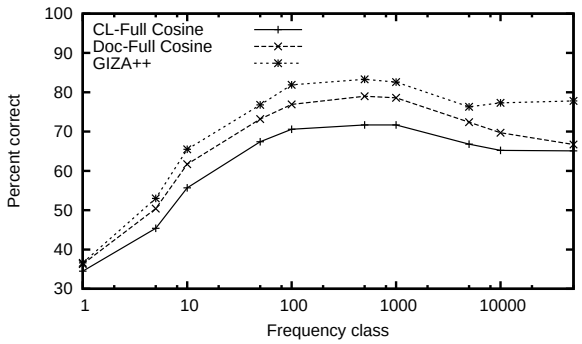

Figure 4: Top performing models compared: syntagmatic (labeled Doc-Full Cosine), paradigmatic (labeled CL-Full Cosine) and statistical (labeled GIZA++).

	Percent correct
Paradigm-Full-Cosine	56.0
Paradigm-Full-MI	52.4
Paradigm-RI-Cosine	55.1
Paradigm-Full-Cosine-CL	**56.2**
Syntagm-Full-Cosine	**61.4**
Syntagm-Full-MI	58.7
Syntagm-RI-Cosine	58.1
GIZA++	64.4 (64.0)
Ensemble	**65.8 (65.3)**

Table 1: Percent correct over all frequency classes, totally 37,316 translations evaluated. Numbers in parenthesis show results when German-French is not included (this direction of translation was used for parameter tuning in the ensemble method).

like to consider when combining the results of the different systems: how confident each system is of its decision (modeled in the S' function below) and how accurate the system has been in the past (modeled in the S'' function below). For each source language term, I look at the top ten translation candidates for each of the three models. The scores for each model are rescaled, so that the scores for the top ten translation candidates for a particular source term sum to one, or, equivalently:

$$S'(x,y) = \frac{S(x,y)}{\sum_{y_i} S(x,y_i)}$$

where x is the source term, y a translation candidate, S the scoring function and S' the rescaled scoring function. I then weight the scores from each model according to how accurately it performed on one direction of translation for one language pair,[9] which I set aside for testing during this particular experiment. The scoring function which is finally used to re-rank the top ten suggestions from the three models looks like this:

$$S''(x,y) = \alpha * S'_a(x,y) + \beta * S'_b(x,y) + \gamma * S'_c(x,y)$$

where α, β and γ are the accuracies of the respective models, normalized so that $\alpha + \beta + \gamma = 1$.[10] Basically, this amounts to the *average combination rule*, which is a standard way of combining multiple classifiers (Tax et al., 2000). The results, displayed in figure 5, show a slight improvement when compared to using the statistical model alone. Finally, table 1 shows the percent correct for each method, regardless of frequency class.

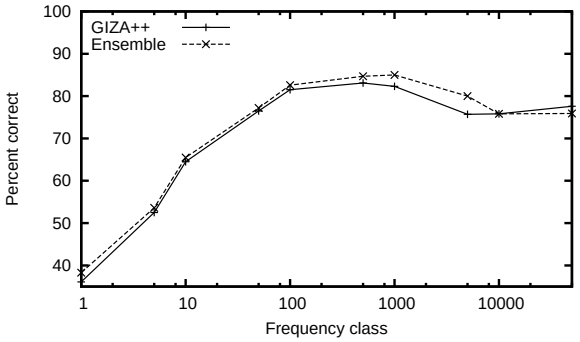

Figure 5: Comparing GIZA++ to the ensemble method.

4 Discussion and future work

Using the non-reduced matrix gives the highest correctness figures, both for the syntagmatic and the paradigmatic models, though the reduced version is trailing closely for the paradigmatic model, as seen in figure 2. There are possible computational benefits of using a reduced representation. However, since both data structures and algorithms designed for working with sparse matrices and vectors exist, one would have to investigate just where the breaking point lies. For the current experiments, using the non-reduced, sparse matrix proved more efficient

[9] I used German to French, to have one Germanic and one Romance language.

[10] This resulted in the following parameters, for the paradigmatic, syntagmatic and statistical models, respectively: $\alpha = 0.313$ $\beta = 0.334$ $\gamma = 0.353$.

both in terms of time and in terms of memory usage, since the reduced matrices have to work with dense representations. It should be noted that when using Random Indexing, the results will vary with the dimensionality of the matrix and the number of non-zero elements used in the random vectors. I used a dimensionality of 1800 and an average of eight non-zero elements (positive and negative), which lies in the range of what is suggested in (Sahlgren, 2006). We note a larger gap in accuracy between the reduced and the full matrix for the syntagmatic models than for the paradigmatic models (0.9% vs. 3.3%). From this we can hypothesize that the reduced syntagmatic model would have performed better using a higher dimensionality, considering that the non-reduced syntagmatic models have a higher dimensionality than the non-reduced paradigmatic models. This is left for future experiments to confirm.

The syntagmatic models consistently outperform the paradigmatic models in these experiments. I am not aware of another study which has directly compared these two approaches on the current task. Further, the cosine measure outperforms the Mutual Information measure in the cases where a direct comparison can be made. This is contrary to what Ribeiro (2000) reported, but the experiments described here have been conducted on a much larger corpus with a larger variability of languages – perhaps this could explain the differences in the results.

Further, the statistical approach clearly outperforms both the paradigmatic and the syntagmatic models. This is again contrary to what Sahlgren and Karlgren (2005) report. However, they claim an accuracy of "something less then 1/3" for the GIZA++ system, which lies far below the 64.4% measured here. The two evaluations can not be directly compared, due to several differences in the methodology of the experiments. The most important difference, which probably by itself explains the vast discrepancy when measuring the performance of GIZA++, is that this study uses texts aligned on a *paragraph* level, whereas Sahlgren and Karlgren used texts aligned on a *document* level. Sahlgren and Karlgren are also studying *words*, not *terms*, which makes their task harder, since they have to pick the correct word out of 40,000 to 70,000 translation candidates, whereas this study typically only has about 3,500 terms as translation candidates. On the other hand, the evaluation applied here is stricter, since only the descriptor in the target language is counted as correct, where Sahlgren and Karlgren also count partial matches in the target language part of a bilingual dictionary as correct.

The correctness for terms occurring only once seems low, at slightly below 40%. Consider, though, that there is no guarantee that the corresponding target language descriptor co-occurs *even once* with these terms. Such cases can arise from e.g., faulty sentence alignment or from the (human) translator choosing to use a different term than the descriptor in the target language translation.

Using the ensemble method described in section 3.5, the results are boosted with 1.3% points. Though the increase is relatively small, the difference is statistically significant beyond the 0.001 level according to McNemar's test. If we use a more lenient evaluation method, counting each result as correct if the corresponding descriptor occurs among the top three translation candidates, GIZA++ achieves 66.9% correct translations on average and the ensemble method reaches 68.6%. Extending this to the top ten candidates, we get 67.2% for GIZA++ and 70.3% for the ensemble method – a difference of 3.1% points. The rather small increases in correctness for GIZA++ using the lenient evaluation methods can most likely be explained by the internal thresholds in the system. Due to these thresholds, GIZA++ most often returns *less* than ten translation candidates for any given source term, which means that the system will not profit as much from using these lenient evaluation schemes.

5 Conclusions

I have compared two distributional models with a statistical method on the task of identifying cross language term equivalents. I have used all directions of translation between four European languages in the evaluation and I have used texts and a thesaurus covering European Union terminology to evaluate the methods. The paradigmatic distributional models were outperformed by the syntagmatic models and the cosine measure worked better than the Mutual Information measure. Both types of distributional models were outperformed by GIZA++, a SMT system. Combining the results of the top per-

forming distributional models with the results of GIZA++ gives a statistically significant increase in accuracy.

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

M. Teresa Cabré Castellví, Rosa Estopà Bagot, and Jordi Vivaldi Palatresi. 2001. Automatic term detection: A review of current systems. In Didier Bourigault, editor, *Recent Advances in Computational Terminology*, chapter 3, pages 53–87. John Benjamins Publishing Company, Philadelphia, PA, USA.

Kenneth Church and William Gale. 1991. Concordances for parallel text. In *Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research*, pages 40–62.

Philipp Cimiano. 2006. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer-Verlag, New York, NY, USA.

Pascale Fung and Kenneth Church. 1994. K-Vec: A new approach for aligning parallel texts. In *Proceedings of COLING 94*, pages 1096–1102. COLING.

Gene H. Golub and Charles F. van Loan. 1996. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, USA, 3 edition.

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Boston, MA, USA.

Hans Hjelm and Christoph Schwarz. 2006. LiSa - morphological analysis for information retrieval. In Stefan Werner, editor, *Proceedings of the 15th NODALIDA conference, Joensuu 2005*, volume 1 of *University of Joensuu electronic publications in linguistics and language technology*. NoDaLiDa, Ling@JoY.

Christian Jacquemin. 2001. *Spotting and Discovering Terms through Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, USA.

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, and Computers*, 28(2):203–208.

Dan Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Charles Kay Ogden and Ivor Armstrong Richards. 1923. *The Meaning of Meaning*. Harcourt, Brace, and World, New York, NY, USA, 8th edition 1946 edition.

António Ribeiro, Gabriel Pereira Lopes, and João Mexia. 2000. Extracting equivalents from aligned parallel texts: Comparison of measures of similarity. In M. C. Monard and J. S. Sichman, editors, *Advances in Artificial Intelligence: International Joint Conference, 7th Ibero-American Conference on AI, 15th Brazilian Symposium on AI, IBERAMIA-SBIA 2000, Atibaia, SP, Brazil, November 2000. Proceedings*, Lecture Notes in Computer Science, pages 339–349. Springer-Verlag, Berlin Heidelberg, Germany.

Magnus Sahlgren and Jussi Karlgren. 2005. Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Natural Language Engineering*, 11(3):327–341.

Magnus Sahlgren. 2006. *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces*. Ph.D. thesis, Stockholm University, Stockholm, Sweden.

Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38.

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

David M. J. Tax, Martin van Breukelen, Robert P. W. Duin, and Josef Kittler. 2000. Combining multiple classifiers by averaging or by multiplying. *Pattern Recognition*, 33(9):1475 – 1485.

Jörg Tiedemann. 2003. *Recycling Translations: Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*. Ph.D. thesis, Uppsala University, Uppsala, Sweden.

Keita Tsuji and Kyo Kageura. 2004. Extracting low-frequency translation pairs from japanese-english bilingual corpora. In Sophia Ananadiou and Pierre Zweigenbaum, editors, *COLING 2004 CompuTerm 2004: 3rd International Workshop on Computational Terminology*, pages 23–30, Geneva, Switzerland. COLING.