A Statistical Word-Level Translation Model for Comparable Corpora

Mona Diab
Linguistics Department & UMIACS
University of Maryland,
College Park, MD 20742
mdiab@umiacs.umd.edu

Steve Finch
West Group
137 Opperman Drive
Eagan, MN
Steve.Finch@ed.ac.uk

Abstract

In this paper, we present a model of statistical word-level mapping for comparable corpora. The approach is based on the assumption that if two terms have close distributional profiles, their corresponding translations' distributional profiles should be close in a comparable corpus. The proposed model is described. A preliminary investigation on intralanguage comparable corpora is laid out. The preliminary results are >92% accurate, suggesting the feasibility of the model. The model needs to undergo some improvements and should be tested cross linguistically before assessing its significance.

Keywords

Word-level mapping, comparable, parallel, Spearman correlation, contingency, gradient descent

1. Introduction

The natural language processing community is in constant need of readily available resources such as corpora, thesauri, bilingual and multilingual lexicons and dictionaries. The acquisition of such resources has proven to be challenging so far, requiring an immense overhead in terms of lexicographers and linguists, especially with the evermore-appealing transition into large scale and very large-scale applications. Many of the existing statistical models for bilingual lexicon creation and machine translation (Brown et al., 1993; Brown et al., 1991; Gale & Church, 1991) depend essentially on the existence of parallel corpora, i.e. translated texts in large amounts. In order to alleviate the expensive investment of human effort, automatic methods have been proposed for the compilation of large amounts of parallel data from the World Wide Web (Resnik, 1999). Yet the problem remains where there are languages that are less represented in electronic forms, let alone in translation into another language. Therefore, it seems natural to be considering alternative data resources such as non-parallel, comparable corpora.

Generally, corpora utilized for statistical translation models take one of two forms: parallel and non-parallel. Parallel corpora are texts existing in translation in two different languages, primarily translated by hand, e.g. the English-French Canadian parliamentary proceedings (Hansards) or the aligned Bible (Resnik et al., 1998). Non-parallel corpora, on the other hand, can be farther subdivided into unrelated and comparable corpora. Unrelated corpora, as the name suggests, are corpora that are of different genres, different sizes or time frames. Comparable corpora are corpora that usually tend to deal with the same genre topics, yet they are usually authored by different people (Oard, 1998). Comparable corpora appear both interlanguage, e.g. New York Times (NYT) in English and Le Monde (LM) in French, and intralanguage, e.g. Wall Street Journal (WSJ) in English and Financial Times (FT) in English. They tend to be of the same size, and covering the same time frame. One can possibly view parallel corpora as a subset of comparable corpora. Interlanguage comparable corpora are a ripe area of investigation in the development of bilingual lexicons and Cross Language Information Retrieval (CLIR) (Peters & Picchi, 1997; Fung & Yee, 1998; Rapp, 1999), and can aid in word-level machine translation, also referred to as Shallow Machine Translation (SMT). Intralanguage comparable corpora, on the other hand, receive less attention, yet they could aid in Monolingual Information Retrieval (MIR) by methods of query expansion, and thesauri construction.

To date, most of the existing statistical models assume the availability of NLP tools such as POS taggers, parsers, morphological analyzers, bilingual lexicons, etc. at least for one of the languages in which the utilized corpora exist in order to bootstrap the system. The aim of this paper is to provide an alternative model where these resources are assumed nonexistent.

We present a statistical word-level mapping model that does not depend on language specific NLP tools[1]. We present a novel technique for statistical word-level translation between comparable corpora in any languages. The method could be applied to any language pair since there is no language specific codification required throughout the translation process. The model can be applied to areas of NLP: SMT, the creation of bilingual word lists therefore aiding in the process of creating bilingual lexicons, thesauri, MIR, and CLIR. In the following section, we give a detailed description of the proposed model, which is validated by a preliminary investigation illustrated in section 3. We discuss the results and related work in sections 4 and 5, respectively. A general discussion of future directions and the conclusion ensue.

2. Approach

The basic intuition is that words that have the same meaning will have similar distributional profiles in language. The approach is an attempt at creating a translation or rather a mapping of tokens that have similar distributional profiles from one corpus to another comparable one. It can be viewed as subjecting one of the corpora to a word-substitution cypher, and attempting to discover that cypher by using statistics of the distribution of tokens within each corpus separately. In principle, we do not have to have the same size corpora in order for the approach to work. Relative distances might converge more quickly with a larger corpus but the technique is relatively insensitive to differences in corpus sizes because we are mapping from one corpus to the other and all the relevant statistics are taken from "within" each corpus rather than "across" them. The approach depends primarily on co-occurrence information of collocate tokens. No morphological or lexical analysis is applied to either corpus during the investigation.

We achieve this by defining a distance metric D between each pair of tokens in each of our corpora, independently, and finding a mapping M of tokens, between the corpora, which preserves the distance mapping as much as possible between these tokens. Suppose we have an English and a French comparable corpus and the token "the" is close to the token "his" in the English corpus. According to the distance metric D, we would want the mapping of the token "le" – which is the mapping M("the") - to be close to the token "lui" – the mapping M("his") - in the French corpus, where closeness is defined quantitatively as in the optimization function in equation [3] below.

Therefore, our goals are: (a) define a distance metric D between tokens which captures similarity between tokens within each of the corpora; and (b) provide an algorithm for deriving a mapping M which captures the substitution cypher between the corpora .The cypher is defined by minimizing disparities between the distances of pairs of tokens under the mapping M. For all pairs of tokens x and y, D(x, y) should be close to D(M(x),M(y))).

In order to measure the distance D, a contingency table is created for the top N most frequent tokens in each of the corpora separately. A fixed sliding window of 2 tokens is used to calculate the co-occurrence frequencies for the most frequent tokens in each of the corpora. A fixed window size is a desirable attribute of the model since it captures semantic similarity rather than syntactic similarity (Manning & Schütze, 302), therefore allowing for a wider range of applications especially cross linguistically, particularly useful for syntactically unrelated languages. The N most highly frequent tokens in a corpus are labeled "focal" terms and extracted. Four vectors are created corresponding to four collocation positions. P2 denotes the collocation of a token and a focal token one token apart in the left context. P1 denotes the collocation of a focal token and its adjacent, to the left, token. Similarly, M1 and M2 define the positions in the right context of a focal token. Each of these positions, P2, P1, M1, and M2, is represented with the same vector of length S, where the dimensions

[1] Except segmenters for languages that do not use space delimiters between words such as Chinese & Arabic

of the vector is defined with the highest most frequent S tokens in a corpus, termed peripheral tokens. Essentially, the S peripheral (pr) tokens were the topmost S tokens from the focal N tokens. Hence, one can view the top S entries of the contingency table, for each of the collocation positions, as a square matrix of size SxS, where the column and row entry labels are the same. The content of each dimension of each of these vectors is defined as the co-occurrence frequency of dimension x with the focal token y in a collocation relation P2, P1, M1, or M2. The N focal tokens constitute the row entries in the contingency table. The columns consist of the four vectors mentioned before, therefore creating a 2-dimensional matrix of Nx4S. Table 1 illustrates this matrix.

Focal token	***P2***				***P1***				***M1***				***M2***			
	pr_1	pr_2	**…**	pr_S	pr_1	pr_2	**…**	pr_S	pr_1	pr_2	**…**	pr_S	pr_1	pr_2	**…**	pr_S
$focal_1$	f_{11}	f_{12}	…	f_{1S}	f_{11}	f_{11}	…	f_{1S}	f_{11}	f_{12}	…	f_{1S}	f_{11}	f_{12}	…	f_{1S}
$focal_2$	f_{21}	f_{22}	…	f_{2S}	f_{21}	f_{22}	…	f_{2S}	f_{21}	f_{22}	…	f_{2S}	f_{21}	f_{22}	…	f_{2S}
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
$focal_S$	f_{S1}	f_{S2}	…	f_{SS}	f_{S1}	f_{S2}		f_{SS}	f_{S1}	f_{S2}	…	f_{SS}	f_{S1}	f_{S2}	…	f_{SS}
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
$focal_N$	f_{N1}	f_{N2}	…	f_{NS}	f_{N1}	f_{N2}	…	f_{NS}	f_{N1}	f_{N2}	…	f_{NS}	f_{N1}	f_{N2}	…	f_{NS}

Table 1: contingency table

In Table 1, *f* denotes the co-occurrence frequency of an element from the top S *pr* tokens with a focal token in one of the tabulated positions P2, P1, M1 or M2, respectively. It is important to note that the values for f_{11}, for instance, will differ depending on the collocation position

A Spearman rank order correlation (R) is used as the distance measure between the focal token vectors from the contingency table. This correlation metric does not assume a linear relation between the elements of the vectors. It measures the monotonic association between the vectors. R is calculated by ranking all the elements in the focal token vector u_p and all the elements in the token vector u_q, separately, and then calculating the regular Pearson product-moment correlation coefficient for the ranks (Ott, 323). In this model, we assume that the data has no ties, therefore the following equation is used to compute R of two focal token vectors:

$$R(\vec{u}_p, \vec{u}_q) = 1 - \frac{6\sum_{i}^{n}(x_i - y_i)}{n(n^2 - 1)} \qquad [1]$$

where $\mathbf{u_p}$ *and* $\mathbf{u_q}$ *are focal token vectors,* $\mathbf{x_i}$ *and* $\mathbf{y_i}$ *are the ranks of the elements in the vectors* $\mathbf{u_p}$ *and* $\mathbf{u_q}$*, respectively, at column* **i** *in the contingency table.* **n** *is the number of columns in the contingency table.*

R as calculated above is a non-parametric correlation measure. Methods of this type tend to have the power of making fewer assumptions regarding the nature of the data, yet they require great amounts of training data (Manning & Schütze, 50).

Once the distance is measured between tokens within a corpus, a cost of mapping C of the pair of tokens vectors' correlation, as in [1], from the source corpus, to a pair of tokens vectors' correlation in a comparable target corpus is calculated as follows:

$$C = R_A(\vec{u}_p, \vec{u}_q) - R_B(M(\vec{u}_p), M(\vec{u}_q)) \qquad [2]$$

where **R** *is defined in [1],* **p** *and* **q** *are defined over the number of focal tokens making up sets* **A** *and* **B***, which are sets of tokens from the source and target corpora, respectively.* **M** *is the mapping function, where it is the mapping of a focal token vector from set A to a focal token vector from set B.*

Goodness is best if *C* is at a minimum, i.e. the distributional profiles of the focal tokens are close enough to each other. The distance mapping between the tokens is based on the following optimization function.

$$DG = \sum (C_{new})^2 - (C_{old})^2 \qquad [3]$$

where ***DG*** *denotes the degree of goodness of the mapping,* ***new*** *denotes a current chosen mapping, as defined in [2] above, of the focal token vectors from set A – source corpus* **-** *onto a pair of focal vector tokens from set B –target corpus* **-** *and* ***old*** *denotes a previous mapping by the same focal token vectors from set A onto a different pair of focal token vectors from set B.*

The algorithm chosen is based on a gradient descent algorithm. Gradient descent algorithms are known for their fast convergence to a solution, although they may reach only a local optimum unless the objective function is known to be convex. The main disadvantage lies in the algorithm's order of computational complexity, usually in the order of O(K*L), where K and L correspond to the number of focal tokens to be mapped from corpus A (K) to corpus B, (L). Effectively, there are two important parts to the algorithm. Firstly, the initial mapping M maps all words onto a virtual token which has distance 3[2] from everything (including itself), except for a few "seed" words (usually punctuation tokens) which are assumed to be common between the two corpora. During descent, the mapping is minimally perturbed (by changing the mapping of a single word) so as to optimize the degree of goodness, as in [3] above, of the new mapping.

3. Preliminary investigation

In order to test the validity of the proposed approach, two comparable corpora are required. Our approach, initially, involves attempting to "translate" between two comparable corpora in the same language. The idea is that if we get a high accuracy in the mappings then we have proven that it is a feasible methodology. We chose a corpus of the economic genre called IAC[3], the content of which is comparable to the WSJ corpus. IAC has 80M words. It is an English corpus. For the investigation, we split the corpus in half creating two comparable corpora of 40M words each, IACA and IACB, respectively.

Each of the corpora went through the same preprocessing phase followed by a token distance calculation phase, independently. The preprocessing phase was done using the Normalized SGML tools[4]. The preprocessing involved a process of SGML marking up, tokenization, counting of the tokens and sorting in a descending order, according to their frequencies in the corpus. No morphological analysis was performed on the data. In this investigation, punctuation marks counted as tokens of interest. The most frequent 2000 (N) tokens (focal) and 150 (S) tokens (pr) were extracted. A contingency table (as in Table 1) of 2000 (N - rows) by 600 (4S - columns) was created. Table 2 shows a sample of a contingency table, which is reproduced here for illustrative purposes.

Focal	***P2***			***P1***			***M1***			***M2***		
token	***,1***	**…**	***rate$_S$***	***,1***	**…**	***rate$_S$***	***,1***	**…**	***rate$_S$***	***,1***	**…**	***rate$_S$***
,1	200	…	120	2	…	150	1	…	310	400	…	309
:	:	:	:	:	:	:	:	:	:	:	:	:
rate$_S$	1000	…	18	150	…	0	965	…	0	800	…	16
rate$_{S+1}$	932	…	31	353	…	0	535	…	0	741	…	64
:	:	:	:	:	:	:	:	:	:	:	:	:
price$_N$	527	…	263	200	…	12	948	…	376	462	…	198

Table 2: Sample contingency table created for illustrative purposes

---

[2] a distance of 3 tokens was empirically decided upon as it yielded the best mapping
[3] IAC was a corpus available to Thomson NLP research labs (proprietary)
[4] URL *http://www.ltg.ed.ac.uk/corpora/nsldoc/nsldoc.html*

The second phase is the token distance calculation, where a Spearman *R* ranked correlation was computed between the focal token row entries in the contingency table, thereby obtaining a measure of similarity between the focal tokens. The correlations are calculated offline and stored in a square matrix NxN, where N is the number of focal elements taken from a corpus.

The next stage is the mapping between the two corpora. Two lists of tokens were created from the two corpora's focal terms, respectively set A, ranging over IACA and set B ranging over IACB. They were mapped to one another using the gradient descent algorithm, where the optimization function is defined in equation [3]. The algorithm was seeded with some of the punctuation marks since they were assumed common to both corpora. Four punctuation marks were used as seeds to bootstrap the descent. Noise is endemic to comparable corpora, e.g. polysemy, so there were cases of many-to-many, many-to-one and one-to-many mappings. A set of mapping experiments was carried out varying the lengths of the token lists A and B –never exceeding 2000 tokens per list - in an attempt at measuring the robustness of the model.

4. Results and Discussion

The preliminary results look extremely promising, especially since none of the traditional tools such as POS taggers, linguistic parsers, or morphological analyzers were used in the process of the investigation. We decided to apply a strong equivalence – identity mapping - for the evaluation phase since we were doing a within-language translation. Therefore, if a token maps onto itself, it was counted as a correct map.

We varied the lengths of the list to check whether there was any deterioration in the performance of the system. The results are illustrated in the following table:

Token list size	**150_A-150_B**	**300_A-300_B**	**300_A-600_B**	**600_A-600_B**	**1000_A-1000_B**	**600_A-1000_B**	**1000_A-600_B**
Accuracy Rate	***98.7%***	***95.3%***	***97%***	***94%***	***92.4%***	***94.6%***	***96.3%***
Sample token mismatches	[._]-[.] [1992] - [1994]	[1993]-[1994] [Company]-[Inc.] [level]-[rate] [3]-[2]	[To]-[of] [level]-[rate]	[1989]-[1990] [employees]-[customers]	[.]-[._] [results]-[prices]	[to]-[of] [.?]- [.!] [performance]-[growth]	[and]-[of]

Table 3: Results Mapping IACA to IACB

In Table 3, the column entries are the number of tokens mapped to one another from the two lists taken from the focal tokens of each corpus. The results, as shown, indicate accuracy rates ranging from 92.4 % to 98.7%. Deterioration in the accuracy rates is noted as lists A and B increase in length suggesting that performance is affected negatively by the size of the lists mapped.

On a closer look at the mismatch list, we observe that the mapping algorithm always mapped tokens onto tokens that have similar meaning or were at least related. For instance, dates were mapped to one another, numbers were mapped to one another, and nouns that are semantically related, such as employees and customers were mapped to each other. This seems to support the idea that this task is useful for the creation of thesauri as well as query expansion for MIR. In fact, one reason for mismatches was the lack of the exact token in both lists. We did not find any instances of a part of speech mismatch, e.g. no instances of a noun mapped to a preposition.

5. Related work

Several successful approaches to use comparable corpora for word to word translation are noted in current literature. In this section, we shall cover those most related to the proposed model. It is worth mentioning that all the relevant work has already been tested on cross language comparable corpora, but, in contrast to our proposal, they all rely on a bilingual dictionary and list of seed words.

(Rapp, 1995) proposes an approach very similar to the model presented here. He builds his model based on the assumption that if two words strongly co-occur – where strength is defined in terms of frequency – then their translations, in comparable and unrelated corpora, will also co-occur with a high frequency. He proposes a model for German-English non-parallel corpora (comprising both comparable and unrelated corpora) which differs essentially in the details of the similarity measure and the word window size, assuming a fixed window size of 11 terms. He uses the city block metric to measure the distance between vectors, or entries in the contingency table. The relevance of this approach to ours lies in the fact that he did not depend on any linguistic tools, e.g. lemmatizers, POS taggers, etc. Later, (Rapp, 1999) reports achieving 72% accuracy rate for German-English word pairs, which is the highest rate to date in statistical word level translation models, for non-parallel corpora interlanguage. The assumption remains the same as in the earlier work by the author, yet he varied the window size for the words to be 4n (12), and he introduced the usage of linguistic tools to the model such as lemmatization, morphological analysis, a bilingual lexicon and seed words. By a close look at the size of each of the utilized corpora 135M and 164M words, respectively, and the bilingual lexicon (>16,000 entries), it is interesting to note the size of the search space, given the window size. The main difference to be noted between our approach and his approach, lies in his usage of linguistic tools, and his eliminatination function words from his investigation. In Rapp's model, the columns in the contingency table express the co-occurrence frequencies of words – if they co-occur within a window size of 12 terms - in German and those obtained from the base lexicon. In our case the co-occurrence frequencies are between the top 2000 frequent tokens in the corpus and the top 150 frequent tokens, in four different collocation positions, as illustrated in Table 1.

(Fung&Yee, 1998) propose an approach based on the vector space model for translating new words in nonparallel, Chinese English comparable corpora. The motivation behind the work is to make use of the easier access to nonparallel resources and arrive at accurate translations for newly encountered words. The basic intuition of their work is that a content word is closely associated with words in its context. They form a vector for a word in terms of its context words, where the vector dimensions are defined by the frequency of occurrence of the context word with the content word in the same sentence, within a corpus. In the similarity measures described in the paper, the magnitude of the data items (term frequencies) is contributing directly to the similarity measure. The frequencies are normalized using the commonly known IR method of Term Frequency (TF) and Inverse Document Frequency (IDF). This contrasts with our model. No assumption is made regarding the distribution of the data, therefore, token frequencies do not contribute directly to the distance measure, rather their ranks with respect to one another, hence, the non-parametric measure of rank correlation. The approach that Fung & Yee propose seems to depend essentially on word pairs from a machine translation system, where these word pairs act as "bridges" between the terms, as well as seeds to bootstrap the word to word translation system. They claim that the association between words and seed words that occur in their context is preserved in comparable corpora, which is consistent with our observations, even when the seed terms are punctuation.

(Peters & Picchi, 1997) propose a method for word-level translation for comparable corpora in Italian and English. The paradigm is slightly different since the model assumes interaction with a user to supply the seed words. It is considered a semi-automatic approach. It relies heavily on the availability of linguistic resources such as bilingual dictionaries and morphological analyzers. They report success for their approach, which is measured in a preliminary investigation for cross-language retrieval.

It is worth noting that the authors of the previous models do not give us a clear indication of how the term equivalency was determined.

One can easily draw a comparison between Latent Semantic Indexing (LSI) and our model. LSI is a variant of the vector space model widely used in IR applications (Dumais et al., 1996). In LSI, one can retrieve relevant documents even if there were no words in common with the query input. LSI hinges upon a significant reduction in the feature space representation, where words that appear in similar contexts would be nearer each other. The method it uses is from linear algebra, Singular Value Decomposition (SVD), in order to discover the associative relationship between the terms. In effect one can view the LSI process as a mapping of both the query and the document into a language independent representation based on term contexts, their co-occurrence frequencies. Our model makes the same claim in representing the top most frequent tokens in terms of their co-occurrence distributional profiles. Hence, our model also reduces the feature space to a set of language independent dimensions. The main difference lies in the choice of the terms on which co-occurrence is measured. In LSI, they are based on training on parallel corpora such as the Canadian Parliamentary (Hansards) collection. The system trains on these parallel documents and produces the LSI space, which consists of terms that are considered identical since they are consistently paired together, and terms that are similar since they are frequently associated with each other, e.g. “not” and “pas”. LSI features a more efficient mapping time than the model we propose. Yet LSI, to date, has been mostly applied where parallel corpora are readily available.

6. Further discussion

The brief comparison to LSI in the previous section allows one to envision a method through which our proposed model can help in both CLIR as well as query expansion in MIR, as mentioned in Section 4. Since terms are represented in terms of their distributional profiles, we have achieved a level of language independence. In case of CLIR, terms from the query can be mapped onto equivalent terms in the target language. The same applies to MIR, since the mapping algorithm allows for a one-to-many mapping. The main disadvantage to our approach lies in the inefficiency of the algorithm, therefore requiring off line processing.

Another drawback of our approach lies in the high sensitivity to the corpus size since there is the urge to gain reliable distinct co-occurrence profiles for each term, hence the cut off point for the number of entries in the contingency table to 2000 elements. Also, the Spearman R rank correlation does not take good account of ties in the data, therefore a Strict Spearman or Gamma coefficient might be utilized to improve performance. Our algorithm needs improvement (over 48hrs on a SPARC 20 for the mapping of a 1000 token list to a 1000 token list). Alternative optimization techniques are being considered, such as simulated annealing or genetic algorithms, which are noted to have more efficient performance. Methods exist, however, in order to reduce the search space in the range (list B), such as applying clustering techniques, so that the comparison will be done only to a representative token.

The results of the current investigation seem promising enough to proceed farther with this approach toward testing the limits of its performance. Future directions include testing the model with a monolingual comparable corpus, e.g. WSJ [42M] and either IACA/B. Furthermore, we would like to test it on parallel and comparable corpora, respectively, for language pairs that are related – English and French – and unrelated language pairs, such as English and Chinese. We would like to investigate the effect of reducing the noise in the data by testing the effect of lemmatization, especially in morphologically rich languages. Automatic evaluation of the results of such experiments is likely to constitute a challenge due to the lack of electronic bilingual dictionaries. Yet one can depend on bilingual speakers' judgements in a carefully designed psycholinguistic study to evaluate system performance.

As mentioned in the introduction, our method serves as an aid in compiling bilingual word lists and monolingual thesauri. It can be viewed as a method of bootstrapping the process of creating bilingual dictionaries, therefore aiding lexicographers in their efforts. Shallow machine translation can benefit from this approach immensely. If this method is coupled with an OCR engine at the input end, it will have solved a resource bottleneck, namely the lack of parallel corpora, in particular for languages that are less likely to be available in an electronic form.

It would be interesting to compare the results of our model once we have results cross linguistically to models of word alignment (Brown et al., 1991; Melamed, 1997). These models get leverage from sentence alignment, which is the reason there is a reliance on parallel corpora, accordingly, using heuristics within the sentence to arrive at word level mapping. Our approach should, in principle, be able to do this mapping with no need for the overhead of sentence alignment.

7. Conclusion

In this paper we have presented a novel approach to statistical word-level mapping between comparable corpora. There is no explicit need for language specific tools for the mapping process. The method is based on the premise that words with similar meaning will have similar distribution in language. The algorithm was presented followed by a preliminary investigation of mapping words intralanguage for a comparable English corpus. The results obtained were very promising, accuracy rates ranging from 92.4% to 98.7%. Future work includes testing with cross language in parallel and comparable corpora and improvements in the order of the algorithm's computational complexity.

Acknowledgements

We would like to acknowledge Thomson research labs where the investigation was carried out. Also we would like to acknowledge Philip Resnik and Douglas Oard for their useful comments and support. The work has been supported in part by DARPA contract N6600197C8540.

References

Brown, P., J. Lai, & R. Mercer (1991). Aligning sentences in parallel corpora. *Proceedings of the 29th Annual Conference of the Association for Computational Linguistics*. pp. 169-176.

Brown, P. F., S.A. Della Pietra, V.J. Della Pietra & R. L. Mercer (1993). The mathematics of machine translation: parameter estimation. *Computational Linguistics*, 19(2): pp. 263-311.

Dumais, Susan T., Thomas K. Landauer & Michael L. Littman (1996). Automatic Cross-Linguistic Information Retrieval using Latent Semantic Indexing. *SIGIR - Workshop on Cross-Linguistic Information Retrieval*, pp. 16-23.

Fung, Pascale & Lo Yuen Yee (1998). An IR Approach for Translating New Words from Nonparallel, Comparable Texts. *Proceedings of the 36th Conference for the Association for Computational Linguistics*, pp. 414 – 420.

Gale, William A. & Kenneth W. Church (1991). A program for aligning sentences in bilingual corpora. *Proceedings of the 29th Annual Conference of the Association for Computational Linguistics.* pp. 177-184.

Manning, Christopher D. & Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT press.

Melamed, I. Dan (1997). Word-to-Word Models of Translational Equivalence. *Proceedings of the 35th Annual Conference of the Association for Computational Linguistics.*

Oard, Douglas W. & Anne R. Diekema (1998). Cross Language Information Retrieval. *Annual Review of Information Science and Technology*, vol. 33, pp. 223-256.

Ott, Lyman (1988). *An Introduction to Statistical Methods and Data Analysis.* Boston, Massachusetts: PWS-KENT Publishing Company.

Peters, C. & E. Picchi (1997). Using Linguistic Tools and Resources in Cross-Language Retrieval. David Hull and Douglas Oard (eds.) *Cross-Language Text and Speech Retrieval Papers from the 1997 AAAI Spring Symposium, Technical Report SS-97-05, AAAI Press,* pp. 179-188

Rapp, Reinhard (1999). Automatic Identification of Word Translations from Unrelated English and German Corpora. *Proceedings of the 37th Annual Conference of the Association for Computational Linguistics.* pp. 519-525.

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

Resnik, Philip, Mari Olsen & Mona Diab (1999). The bible as a parallel corpus: annotating the "book of 2000 tongues". *Computers and the Humanities*, vol. 33: pp. 129-153.

Resnik, Philip (1999). Mining the Web for Bilingual Text. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 527-534.

Sheridan, Paraic & Peter Schauble (1997). Cross-Language Multi-Media Information Retrieval. In 3rd DELOS Workshop; Cross-Language Information Retrieval, number 97-W003 in ERCIM Workshop Proceedings. European Research Consortium for Informatics and Mathematics, March.

Filename:	RIAO.rtf
Directory:	F:\denise\amy\treports\MONA\MONA
Template:	C:\Program Files\Microsoft Office\Office\Normal.dot
Title:	A statistical Translational Model for Comparable corpora
Subject:	
Author:	mona diab
Keywords:	
Comments:	
Creation Date:	3/26/00 1:04 PM
Change Number:	2
Last Saved On:	3/26/00 1:04 PM
Last Saved By:	mona diab
Total Editing Time:	1 Minute
Last Printed On:	6/15/00 5:40 PM
As of Last Complete Printing	
Number of Pages:	9
Number of Words:	4,597 (approx.)
Number of Characters:	26,208 (approx.)