An IR Approach for Translating New Words from Nonparallel, Comparable Texts

Pascale Fung and **Lo Yuen Yee**
HKUST
Human Language Technology Center
Department of Electrical and Electronic Engineering
University of Science and Technology
Clear Water Bay, Hong Kong
{pascale,eeyy}@ee.ust.hk

1 Introduction

In recent years, there is a phenomenal growth in the amount of online text material available from the greatest information repository known as the World Wide Web. Various traditional information retrieval(IR) techniques combined with natural language processing(NLP) techniques have been re-targeted to enable efficient access of the WWW—search engines, indexing, relevance feedback, query term and keyword weighting, document analysis, document classification, etc. Most of these techniques aim at efficient online search for information already on the Web.

Meanwhile, the corpus linguistic community regards the WWW as a vast potential of corpus resources. It is now possible to download a large amount of texts with automatic tools when one needs to compute, for example, a list of synonyms; or download domain-specific monolingual texts by specifying a keyword to the search engine, and then use this text to extract domain-specific terms. It remains to be seen how we can also make use of the multilingual texts as NLP resources.

In the years since the appearance of the first papers on using statistical models for bilingual lexicon compilation and machine translation(Brown et al., 1993; Brown et al., 1991; Gale and Church, 1993; Church, 1993; Simard et al., 1992), large amount of human effort and time has been invested in collecting parallel corpora of translated texts. Our goal is to alleviate this effort and enlarge the scope of corpus resources by looking into monolingual, comparable texts. This type of texts are known as *nonparallel* corpora. Such nonparallel, monolingual texts should be much more prevalent than parallel texts. However, previous attempts at using nonparallel corpora for terminology translation were constrained by the inadequate availability of same-domain, comparable texts *in electronic form*. The type of nonparallel texts obtained from the LDC or university libraries were often restricted, and were usually out-of-date as soon as they became available. For new word translation, the timeliness of corpus resources is a prerequisite, so is the continuous and automatic availability of nonparallel, comparable texts in electronic form. Data collection effort should not inhibit the actual translation effort. Fortunately, nowadays the World Wide Web provides us with a daily increase of fresh, up-to-date multilingual material, together with the archived versions, all easily downloadable by software tools running in the background. It is possible to specify the URL of the online site of a newspaper, and the start and end dates, and automatically download all the daily newspaper materials between those dates.

In this paper, we describe a new method which combines IR and NLP techniques to extract new word translation from automatically downloaded English-Chinese nonparallel newspaper texts.

2 Encountering new words

To improve the performance of a machine translation system, it is often necessary to update its bilingual lexicon, either by human lexicographers or statistical methods using large corpora. Up until recently, statistical bilingual lexicon compilation relies largely on parallel corpora. This is an undesirable constraint at times. In using a broad-coverage English-Chinese MT system to translate some text recently, we discovered that it is unable to translate 流感/*liougan* which occurs very frequently in the text. Other words which the system cannot find in its 20,000-entry lexicon include proper names

such as the Taiwanese president *Lee Teng-Hui*, and the Hong Kong Chief Executive *Tung Chee-Hwa*. To our disappointment, we cannot locate any parallel texts which include such words since they only start to appear frequently in recent months.

A quick search on the Web turned up archives of multiple local newspapers in English and Chinese. Our challenge is to find the translation of 流感/*liougan* and other words from this online nonparallel, comparable corpus of newspaper materials. We choose to use issues of the English newspaper *Hong Kong Standard* and the Chinese newspaper *Mingpao*, from Dec.12,97 to Dec.31,97, as our corpus. The English text contains about 3 Mb of text whereas the Chinese text contains 8.8 Mb of 2 byte character texts. So both texts are comparable in size. Since they are both local mainstream newspapers, it is reasonable to assume that their contents are comparable as well.

3 流感/*liougan* is associated with *flu* but not with *Africa*

Unlike in parallel texts, the position of a word in a text does not give us information about its translation in the other language. (Rapp, 1995; Fung and McKeown, 1997) suggest that a content word is closely associated with some words in its context. As a tutorial example, we postulate that the words which appear in the context of 流感/*liougan* should be *similar* to the words appearing in the context of its English translation, *flu*. We can form a vector space model of a word in terms of its context word indices, similar to the vector space model of a text in terms of its constituent word indices (Salton and Buckley, 1988; Salton and Yang, 1973; Croft, 1984; Turtle and Croft, 1992; Bookstein, 1983; Korfhage, 1995; Jones, 1979).

The value of the i-th dimension of a word vector W is f if the i-th word in the lexicon appears f times in the *same sentences as* W.

Left columns in Table 1 and Table 2 show the list of content words which appear most frequently in the context of *flu* and *Africa* respectively. The right column shows those which occur most frequently in the context of 流感. We can see that the context of 流感 is more similar to that of *flu* than to that of *Africa*.

Table 1: 流感 and *flu* have similar contexts

English	Freq.	Chinese	Freq.
bird	170	病毒 (virus)	147
virus	26	市民 (citizen)	90
spread	17	香港 (Hong Kong)	84
people	17	感染 (infection)	69
government	13	證實 (confirmed)	62
avian	11	表示 (show)	62
scare	10	發現 (discover)	56
deadly	10	昨日 (yesterday)	54
new	10	病人 (patient)	53
suspected	9	懷疑 (suspected)	50
chickens	9	醫生 (doctor)	49
spreading	8	染上 (infected)	47
prevent	8	醫院 (hospital)	44
crisis	8	沒有 (no)	42
health	8	政府 (government)	41
symptoms	7	事件 (event)	40

Table 2: 流感 and *Africa* have different contexts

English	Freq.	Chinese	Freq.
South	109	病毒 (virus)	147
African	32	市民 (citizen)	90
China	20	香港 (Hong Kong)	84
ties	15	感染 (infection)	69
diplomatic	14	證實 (confirmed)	62
Taiwan	12	表示 (show)	62
relations	9	發現 (discover)	56
Test	9	昨日 (yesterday)	54
Mandela	8	病人 (patient)	53
Taipei	7	懷疑 (suspected)	50
Africans	7	醫生 (doctor)	49
January	7	染上 (infected)	47
visit	6	醫院 (hospital)	44
tense	6	沒有 (no)	42
survived	6	政府 (government)	41
Beijing	6	事件 (event)	40

4 Bilingual lexicon as seed words

So the first clue to the similarity between a word and its translation number of common words in their contexts. In a bilingual corpus, the "common word" is actually a bilingual word pair. We use the lexicon of the MT system to "bridge" all bilingual word pairs in the corpora. These word pairs are used as **seed words**.

We found that the contexts of *flu* and 流感/*liougan* share 233 "common" context words, whereas the contexts of *Africa* and 流感/*liougan* share only 121 common words, even though the context of *flu* has 491 unique words and the context of *Africa* has 328 words.

In the vector space model, $W[flu]$ and $W[liougan]$ has 233 overlapping dimensions, whereas there are 121 overlapping dimensions between $W[flu]$ and $W[Africa]$.

5 Using TF/IDF of contextual seed words

The *flu* example illustrates that the actual ranking of the context word frequencies provides a second clue to the similarity between a bilingual word pair. For example, *virus* ranks very high for both *flu* and 流感/*liougan* and is a strong "bridge" between this bilingual word pair. This leads us to use the term frequency(TF) measure. The TF of a context word is defined as the frequency of the word *in the context of* W. (e.g. TF of *virus* in *flu* is 26, in 流感 is 147).

However, the TF of a word is not independent of its general usage frequency. In an extreme case, the function word *the* appears most frequently in English texts and would have the highest TF in the context of any W. In our HK-Standard/Mingpao corpus, *Hong Kong* is the most frequent content word which appears everywhere. So in the *flu* example, we would like to reduce the significance of *Hong Kong*'s TF while keeping that of *virus*. A common way to account for this difference is by using the inverse document frequency(IDF). Among the variants of IDF, we choose the following representation from (Jones, 1979):

$$\begin{aligned} \text{IDF} &= \log \frac{\text{maxn}}{n_i} + 1 \\ \text{where maxn} &= \text{the maximum frequency of any word in the corpus} \\ n_i &= \text{the total number of occurrences of word } i \text{ in the corpus} \end{aligned}$$

The IDF of *virus* is 1.81 and that of *Hong Kong* is 1.23 in the English text. The IDF of 流感 is 1.92 and that of *Hong Kong* is 0.83 in Chinese. So in both cases, *virus* is a stronger "bridge" for 流感/*liougan* than *Hong Kong*.

Hence, for every context seed word i, we assign a **word weighting factor** (Salton and Buckley, 1988) $w_i = TF_{iW} \times IDF_i$ where TF_{iW} is the TF of word i in the context of word W. The updated vector space model of word W has w_i in its i-th dimension.

The ranking of the 20 words in the contexts of 流感/*liougan* is rearranged by this weighting factor as shown in Table3.

Table 3: *virus* is a stronger bridge than *Hong Kong*

bird	259.97	病毒 (virus)	282.70
spread	51.41	感染 (infection)	187.50
virus	47.07	市民 (citizens)	163.49
avian	43.41	證實 (confirmed)	161.89
scare	36.65	染上 (infected)	158.43
deadly	35.15	病人 (patient)	132.14
spreading	30.49	懷疑 (suspected)	123.08
suspected	28.83	醫生 (doctor)	108.54
symptoms	28.43	醫院 (hospital)	102.73
prevent	26.93	發現 (discover)	98.09
people	23.09	事件 (event)	83.75
crisis	22.72	香港 (Hong Kong)	69.68
health	21.97	昨日 (yesterday)	66.84
new	17.80	可能 (possible)	60.20
government	16.04	表示 (no)	59.76
chickens	15.12	情況 (government)	59.41

6 Ranking translation candidates

Next, a ranking algorithm is needed to match the unknown word vectors to their counterparts in the other language. A ranking algorithm selects the best target language candidate for a source language word according to direct comparison of some similarity measures (Frakes and Baeza-Yates, 1992).

We modify the similarity measure proposed by (Salton and Buckley, 1988) into the following $S0$:

$$\begin{aligned} S0(W_c, W_e) &= \frac{\Sigma_{i=1}^{t}(w_{ic} \times w_{ie})}{\sqrt{\Sigma_{i=1}^{t} {w_{ic}}^2 \times \Sigma_{i=1}^{t} {w_{ie}}^2}} \\ \text{where } w_{ic} &= TF_{ic} \\ w_{ie} &= TF_{ie} \end{aligned}$$

Variants of similarity measures such as the above have been used extensively in the IR community (Frakes and Baeza-Yates, 1992). They are mostly based on the Cosine Measure of two vectors. For different tasks, the weighting factor might vary. For example, if we add the IDF into the weighting factor, we get the following measure $S1$:

$$\begin{aligned} S1(W_c, W_e) &= \frac{\Sigma_{i=1}^{t}(w_{ic} \times w_{ie})}{\sqrt{\Sigma_{i=1}^{t} {w_{ic}}^2 \times \Sigma_{i=1}^{t} {w_{ie}}^2}} \\ \text{where } w_{ic} &= TF_{ic} \times IDF_i \\ w_{ie} &= TF_{ie} \times IDF_i \end{aligned}$$

In addition, the Dice and Jaccard coefficients are also suitable similarity measures for document comparison (Frakes and Baeza-Yates, 1992). We also implement the Dice coefficient into similarity measure $S2$:

$$\begin{aligned} S2(W_c, W_e) &= \frac{2\Sigma_{i=1}^{t}(w_{ic} \times w_{ie})}{\Sigma_{i=1}^{t} {w_{ic}}^2 + \Sigma_{i=1}^{t} {w_{ie}}^2} \\ \text{where } w_{ic} &= TF_{ic} \times IDF_i \\ w_{ic} &= TF_{ie} \times IDF_i \end{aligned}$$

$S1$ is often used in comparing a short query with a document text, whereas $S2$ is used in comparing two document texts. Reasoning that our objective falls somewhere in between—we are comparing segments of a document, we also multiply the above two measures into a third similarity measure $S3$.

7 Confidence on seed word pairs

In using bilingual seed words such as 病毒/*virus* as "bridges" for terminology translation, the quality of the bilingual seed lexicon naturally affects the system output. In the case of European language pairs such as French-English, we can envision using words sharing common cognates as these "bridges". Most importantly, we can assume that the word boundaries are similar in French and English. However, the situation is messier with English and Chinese. First, segmentation of the Chinese text into words already introduces some ambiguity of the seed word identities. Secondly, English-Chinese translations are complicated by the fact that the two languages share very little stemming properties, or part-of-speech set, or word order. This property causes every English word to have many Chinese translations and vice versa. In a source-target language translation scenario, the translated text can be "rearranged" and cleaned up by a monolingual language model in the target language. However, the lexicon is not very reliable in establishing "bridges" between non-parallel English-Chinese texts. To compensate for this ambiguity in the seed lexicon, we introduce a **confidence weighting** to each bilingual word pair used as seed words. If a word i_e is the $k-th$ candidate for word i_c, then $w_{i_{te}} = w_{i_{te}}/k_i$.

The similarity scores then become $S4$ and $S5$ and $S6 = S4 \times S5$:

$$\begin{aligned} S4(W_c, W_e) &= \frac{\Sigma_{i=1}^{t}(w_{ic} \times w_{ie})/k_i}{\sqrt{\Sigma_{i=1}^{t} {w_{ic}}^2 \times \Sigma_{i=1}^{t} {w_{ie}}^2}} \\ \text{where } w_{ic} &= TF_{ic} \times IDF_i \\ w_{ie} &= TF_{ie} \times IDF_i \\ S5(W_c, W_e) &= \frac{2\Sigma_{i=1}^{t}(w_{ic} \times w_{ie})/k_i}{\Sigma_{i=1}^{t} {w_{ic}}^2 + \Sigma_{i=1}^{t} {w_{ie}}^2} \\ \text{where } w_{ic} &= TF_{ic} \times IDF_i \\ w_{ie} &= TF_{ic} \times IDF_i \end{aligned}$$

We also experiment with other combinations of the similarity scores such as $S7 = S0 \times S5$. All similarity measures $S3 - S7$ are used in the experiment for finding a translation for 流感.

8 Results

In order to apply the above algorithm to find the translation for 流感/*liougan* from the HKStandard/Mingpao corpus, we first use a script to select the 118 English content words which are not in the lexicon as possible candidates. Using similarity measures $S3-S7$, the highest ranking candidates of 流感 are shown in Table 6. $S6$ and $S7$ appear to be the best similarity measures.

We then test the algorithm with $S7$ on more Chinese words which are not found in the lexicon but which occur frequently enough in the Mingpao texts. A statistical new word extraction tool can be used to find these words. The unknown Chinese words and their English counterparts, as well as the occurrence frequencies of these words in HKStandard/Mingpao are shown in Table 4. Frequency numbers with a * indicates that this word does not occur frequent enough to be found. Chinese words with a * indicates that it is a word with segmentation and translation ambiguities. For example, 林 *(Lam)* could be a family name, or part of another word meaning *forest.* When it is used as a family name, it could be transliterated into *Lam* in Cantonese or *Lin* in Mandarin.

Disregarding all entries with a * in the above table, we apply the algorithm to the rest of the Chinese unknown words and the 118 English unknown words from HKStandard. The output is ranked by the similarity scores. The highest ranking translated pairs are shown in Table 5.

The only Chinese unknown words which are not correctly translated in the above list are 農

Table 4: Unknown words which occur often

Freq.	Chinese	Freq.	English
59	銅鑼灣 (Causeway)	37*	Causeway
1965	周 (Chau)*	49	Chau
481	建華 (Chee-hwa)	77	Chee-hwa
115	赤 (Chek)*	28	Chek
164	戴安娜 (Diana)	100	Diana
3164	方 (Fong)*	32	Fong
2274	香港 (HONG)	60	HONG
1128	黃 (Huang)*	30	Huang
477	葉 (Ip)*	32	Ip
1404	林 (Lam)*	175	Lam
687	劉 (Lau)*	111	Lau
324	鴨 (Lei)	30	Lei
967	梁 (Leung)	145	Leung
312	農曆 (Lunar)	36	Lunar
164	首相 (Minister)	197	Minister
949	個人 (Personal)	8*	Personal
56	色情 (Pornography)	13*	Pornography
493	家禽 (Poultry)	57	Poultry
1027	主席 (President)	239	President
946	錢 (Qian)*	62	Qian
154	其琛 (Qichen)	28*	Qichen
824	特區 (SAR)	142	SAR
325	譚 (Tam)*	154	Tam
281	唐 (Tang)	80	Tang
307	登輝 (Teng-hui)	37	Teng-hui
350	屯 (Tuen)	76	Tuen
1052	董 (Tung)	274	Tung
79	范 (Versace)*	74	Versace
107	葉利欽 (Yeltsin)	100	Yeltsin
112	珠海 (Zhuhai)	76	Zhuhai
1171	流感 (flu)	491	flu

曆/*Lunar* and 葉利欽/*Yeltsin* [1]. *Tung/Chee-Hwa* is a pair of collocates which is actually the full name of the Chief Executive. *Poultry* in Chinese is closely related to *flu* because the Chinese name for *bird flu* is *poultry flu*. In fact, almost all unambiguous Chinese new words find their translations in the first 100 of the ranked list. Six of the Chinese words have correct translation as their first candidate.

9 Related work

Using vector space model and similarity measures for ranking is a common approach in IR for query/text and text/text comparisons (Salton and Buckley, 1988; Salton and Yang, 1973; Croft, 1984; Turtle and Croft, 1992; Bookstein, 1983; Korfhage, 1995; Jones, 1979). This approach has also been used by (Dagan and Itai, 1994; Gale et al., 1992; Shütze, 1992; Gale et al., 1993; Yarowsky, 1995; Gale and Church, 1994) for sense disambiguation between multiple usages of the same word. Some of the early statistical terminology translation methods are (Brown et al., 1993; Wu and Xia, 1994; Dagan and Church, 1994; Gale and Church, 1991; Kupiec, 1993; Smadja et al., 1996; Kay and Röscheisen, 1993; Fung and Church, 1994; Fung, 1995b). These algorithms all require parallel, translated texts as input. Attempts at exploring nonparallel corpora for terminology translation are very few (Rapp, 1995; Fung, 1995a; Fung and McKeown, 1997). Among these, (Rapp, 1995) proposes that the association between a word and its close collocate is preserved in any language. and (Fung and McKeown, 1997) suggests that the associations between a word and many seed words are also preserved in another language. In this paper,

[1] *Lunar* is not an unknown word in English, *Yeltsin* finds its translation in the 4-th candidate.

Table 5: Some Chinese unknown word translation output

score	English	Chinese
0.008421	Teng-hui	登輝 (Teng-hui)
0.007895	SAR	特區 (SAR)
0.007669	flu	流感 (flu)
0.007588	Lei	鴨 (Lei)
0.007283	poultry	家禽 (Poultry)
0.006812	SAR	建華 (Chee-hwa)
0.006430	hijack	登輝 (Teng-hui)
0.006218	poultry	特區 (SAR)
0.005921	Tung	建華 (Chee-hwa)
0.005527	Diaoyu	登輝 (Teng-hui)
0.005335	PrimeMinister	登輝 (Teng-hui)
0.005335	President	登輝 (Teng-hui)
0.005221	China	林 (Lam)
0.004731	Lien	登輝 (Teng-hui)
0.004470	poultry	建華 (Chee-hwa)
0.004275	China	登輝 (Teng-hui)
0.003878	flu	鴨 (Lei)
0.003859	PrimeMinister	建華 (Chee-hwa)
0.003859	President	建華 (Chee-hwa)
0.003784	poultry	梁 (Leung)
0.003686	Kalkanov	珠海 (Zhuhai)
0.003550	poultry	鴨 (Lei)
0.003519	SAR	葉利欽 (Yeltsin)
0.003481	Zhuhai	建華 (Chee-hwa)
0.003407	PrimeMinister	林 (Lam)
0.003407	President	林 (Lam)
0.003338	flu	家禽 (Poultry)
0.003324	apologise	登輝 (Teng-hui)
0.003250	DPP	登輝 (Teng-hui)
0.003206	Tang	唐 (Tang)
0.003202	Tung	梁 (Leung)
0.003040	Leung	梁 (Leung)
0.003033	China	特區 (SAR)
0.002888	Zhuhai	農曆 (Lunar)
0.002886	Tung	董 (Tung)

we have demonstrated that the associations between a word and its context seed words are well-preserved in nonparallel, comparable texts of different languages.

10 Discussions

Our algorithm is the first to have generated a collocation bilingual lexicon, albeit small, from a nonparallel, comparable corpus. We have shown that the algorithm has good precision, but the recall is low due to the difficulty in extracting unambiguous Chinese and English words.

Better results can be obtained when the following changes are made:

- improve seed word lexicon reliability by stemming and POS tagging on both English and Chinese texts;
- improve Chinese segmentation by using a larger monolingual Chinese lexicon;
- use larger corpus to generate more unknown words and their candidates by statistical methods;

We will test the precision and recall of the algorithm on a larger set of unknown words.

11 Conclusions

We have devised an algorithm using **context seed word TF/IDF** for extracting bilingual lexicon from **nonparallel, comparable corpus** in English-Chinese. This algorithm takes into account the reliability of bilingual seed words and is language independent. This algorithm can be applied to other language pairs such as English-French or English-German. In these cases, since the languages are more similar linguistically and the seed word lexicon is more reliable, the algorithm should yield better results. This algorithm can also be applied in an iterative fashion where high-ranking bilingual word pairs can be added to the seed word list, which in turn can yield more new bilingual word pairs.

Table 6: English words most similar to 流感/*liougan*

S0		
0.181114	Lei	流感
0.088879	flu	流感
0.085886	Tang	流感
0.081411	Ap	流感
S4		
0.120879	flu	流感
0.097577	Lei	流感
0.068657	Beijing	流感
0.065833	poultry	流感
S5		
0.086287	flu	流感
0.040090	China	流感
0.028157	poultry	流感
0.024500	Beijing	流感
S6		
0.010430	flu	流感
0.001854	poultry	流感
0.001840	China	流感
0.001682	Beijing	流感
S7		
0.007669	flu	流感
0.001956	poultry	流感
0.001669	China	流感
0.001391	Beijing	流感

References

A. Bookstein. 1983. Explanation and generalization of vector models in information retrieval. In *Proceedings of the 6th Annual International Conference on Research and Development in Information Retrieval*, pages 118–132.

P. Brown, J. Lai, and R. Mercer. 1991. Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Conference of the Association for Computational Linguistics.*

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

Kenneth Church. 1993. Char_align: A program for aligning parallel texts at the character level. In *Proceedings of the 31st Annual Conference of the Association for Computational Linguistics*, pages 1–8, Columbus, Ohio, June.

W. Bruce Croft. 1984. A comparison of the cosine correlation and the modified probabilistic model. In *Information Technology*, volume 3, pages 113–114.

Ido Dagan and Kenneth W. Church. 1994. Termight: Identifying and translating technical terminology. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, pages 34–40, Stuttgart, Germany, October.

Ido Dagan and Alon Itai. 1994. Word sense disambiguation using a second language monolingual corpus. In *Computational Linguistics*, pages 564–596.

William B. Frakes and Ricardo Baeza-Yates, editors. 1992. *Information Retrieval: Data structures & Algorithms.* Prentice-Hall.

Pascale Fung and Kenneth Church. 1994. Kvec: A new approach for aligning parallel texts. In *Proceedings of COLING 94*, pages 1096–1102, Kyoto, Japan, August.

Pascale Fung and Kathleen McKeown. 1997. Finding terminology translations from non-parallel corpora. In *The 5th Annual Workshop on Very Large Corpora*, pages 192–202, Hong Kong, Aug.

Pascale Fung and Dekai Wu. 1994. Statistical augmentation of a Chinese machine-readable dictionary. In *Proceedings of the Second Annual Workshop on Very Large Corpora*, pages 69–85, Kyoto, Japan, June.

Pascale Fung. 1995a. Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In *Proceedings of the Third Annual Workshop on Very Large Corpora*, pages 173–183, Boston, Massachusettes, June.

Pascale Fung. 1995b. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *Proceedings of the 33rd Annual Conference of the Association for Computational Linguistics*, pages 236–233, Boston, Massachusettes, June.

William Gale and Kenneth Church. 1991. Identifying word correspondences in parallel text. In *Proceedings of the Fourth Darpa Workshop on Speech and Natural Language*, Asilomar.

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

William A. Gale and Kenneth W. Church. 1994. Discrimination decisions in 100,000 dimensional spaces. *Current Issues in Computational Linguisitcs: In honour of Don Walker*, pages 429–550.

W. Gale, K. Church, and D. Yarowsky. 1992. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proceedings of the 30th Conference of the Association for Computational Linguistics.* Association for Computational Linguistics.

W. Gale, K. Church, and D. Yarowsky. 1993. A method for disambiguating word senses in a large corpus. In *Computers and Humanities*, volume 26, pages 415–439.

K. Sparck Jones. 1979. Experiments in relevance weighting of search terms. In *Information Processing and Management*, pages 133–144.

Martin Kay and Martin Röscheisen. 1993. Text-Translation alignment. *Computational Linguistics*, 19(1):121–142.

Robert Korfhage. 1995. Some thoughts on similarity measures. In *The SIGIR Forum*, volume 29, page 8.

Julian Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st Annual Conference of the Association for Computational Linguistics*, pages 17–22, Columbus, Ohio, June.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 35th Conference of the Association of Computational Linguistics, student session*, pages 321–322, Boston, Mass.

G. Salton and C. Buckley. 1988. Term-weighting approaches in automatic text retrieval. In *Information Processing and Management*, pages 513–523.

G. Salton and C. Yang. 1973. *On the specification of term values in automatic indexing*, volume 29.

Hinrich Shütze. 1992. Dimensions of meaning. In *Proceedings of Supercomputing '92.*

M. Simard, G Foster, and P. Isabelle. 1992. Using cognates to align sentences in bilingual corpora. In *Proceedings of the Forth International Conference on Theoretical and Methodological Issues in Machine Translation*, Montreal, Canada.

Frank Smadja, Kathleen McKeown, and Vasileios Hatzsivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 21(4):1–38.

Howard R. Turtle and W. Bruce Croft. 1992. A comparison of text retrieval methods. In *The Computer Journal*, volume 35, pages 279–290.

Dekai Wu and Xuanyin Xia. 1994. Learning an English-Chinese lexicon from a parallel corpus. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, pages 206–213, Columbia, Maryland, October.

D. Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Conference of the Association for Computational Linguistics*, pages 189–196. Association for Computational Linguistics.