Preliminary Examination Report: Bilingual Terminology Translation in Scientific Literature using Multilingual Structural Clues

Benjamin King
Department of Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, MI
benjaminking@umich.edu

Abstract

In this report, we consider the problem of terminology translation in non-parallel scientific literature. We hypothesize that multilingual scientific literature has significant structure, both in the front matter and citations, that can be exploited for improved translation. After testing the well-known context vector method for translation from comparable corpora, we then propose an improvement which is able to exploit certain nearly-parallel segments of text present in the structure of multi-lingual scientific literature. Experimental results indicate that this method gives a substantial improvement in the ranking and accuracy of the system.

1 Introduction

Most machine translation relies on parallel corpora (Lopez, 2008), but in many domains, it can be difficult to acquire a volume of parallel text sufficient to train such a translation system satisfactorily. One such domain is that of scientific literature, and the problem is exacerbated particularly when one considers that much of scientific terminology is rarely used outside of a specific domain, and may even be not be common within the domain.

Multilingual scientific literature can be thought of as forming comparable corpora, multiple non-parallel text collections which share the same domain or topic. Though such corpora may lack parallelism, which is crucial for most machine translation, text of this type has nonetheless been proven useful in bilingual translation mining. (See Related Work)

Scientific literature by nature also has a great degree of structure, both in the front matter necessary for publication, such as a title, an abstract, authors, and keywords, and in the citation structure, including both the references and the citing sentences. We believe that this structure can provide clues about the best translation for a term.

In this paper we evaluate one well-known existing technique for terminology translation from comparable corpora in a realistic scenario and propose an extension that exploits the structure and multilingualism of papers by aligning "nearly-parallel" segments and is shown to greatly improve the accuracy.

2 Related Work

Broadly, there have been three major approaches to terminology translation: (1) from parallel corpora, (2) from comparable corpora, and (3) from mostly monolingual corpora.

Terminology translation from parallel corpora was one of the early problems considered as statistical translation emerged (Dagan and Church, 1994; Smadja et al., 1996). These studies consistently demonstrated the need for larger corpora in order to improve performance.

The earliest work on translation from comparable corpora was done by Pascale Fung (Fung and McKeown, 1997; Fung and Yee, 1998). Her earlier work with McKeown dealt specifically with terminology

Language	# of documents	# of words
English	18,041	107,681,469
Spanish	1,135	4,033,316

Table 1: Statistics for the English and Spanish corpora.

translation and used a "Word Relation Matrix" to equate pairs of words that appear in similar contexts. The contexts between the two languages were normalized by using a list of seed words, whose translations in the other language were known. The accuracy of the top suggested term was only about 30% (Fung and McKeown, 1997).

In Fung's next paper, she refined the technique by creating a context vector for every word, which represented the average context in which a word appeared over all the documents in each corpus. They also demonstrated the plausibility of the hypothesis that the contexts of words in a domain are preserved across languages, specifically English and Chinese (Fung and Yee, 1998).

Munteanu and Marcu have also done relevant research in extracting parallel sentences from comparable corpora (Munteanu and Marcu, 2006). Using a Maximum Entropy classifier, they found pairs of sentences that were nearly parallel, to the extent that they could be used for traditional statistical machine translation.

A third approach works on primarily monolingual corpora. This is applicable for languages in which neologisms are typically presented along with an English translation. Similar approaches were proposed by both (Wu et al., 2004) and (Lin et al., 2008). Both mine sentences containing English parentheticals, making the assumption that the translation of the parenthetical text appears in the same sentence and is more likely to appear near the parenthetical. The use of these structural clues directly influenced our use of multilingual structural patterns in this paper.

Our work most closely follows from the more recent work of Morin, et al., who used a version of the context vector approach modified to handle both single words and multi-word terms. (Morin et al., 2007) They obtained accuracy similar to Fung, et al. in the case of single word terms and slightly lower accuracy in the case of multi-word terms, which were not able be translated using previous context vector approaches. That the accuracy is lower for multi-word terms is blamed on the lower frequency of multi-word terms as compared to single-word terms.

3 Lexical Resources

3.1 Corpora

Our two comparable corpora consist of papers in Spanish and English on the topic of Natural Language Processing. The English corpus is the ACL Anthology (Bird et al., 2008). The Spanish corpus is a collection of 1,135 Spanish language papers available from the website of the Spanish journal "Sociedad Española para el Procesamiento del Lenguaje Natural" (SEPLN)[1]. Table 1 compares the two corpora.

3.2 Bilingual Dictionary

To construct a Spanish-English dictionary, we combined four large publicly available dictionaries[2]: OmegaWiki, Wiktionary, the dicts.info Universal Dictionary, and a dictionary from the Apertium Project (Tyers et al., 2010).

4 Methods

We produce and ordered list of translation candidates from the context vector method and provide this list as input to a reranker that uses structural clues from documents in the two corpora to produce another

[1] http://www.sepln.org/

[2] The first three are available from http://www.dicts.info/

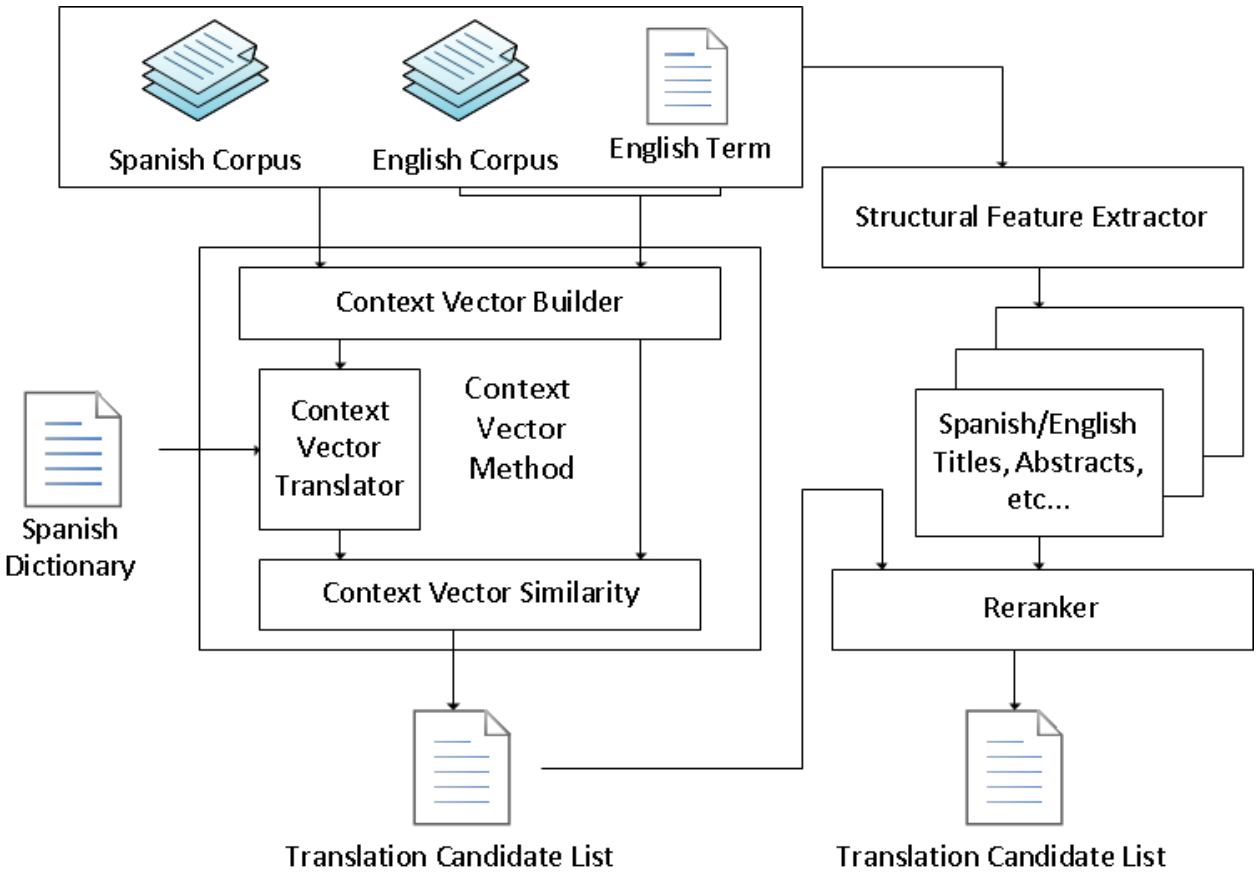

Figure 1: A diagram of the system built in this paper, starting with the lexical resources and ending with the translation candidate lists produced by each of the context vector method and the reranking method.

translation candidate list. (See Figure 1 for a visual representation of this process.) We then compare these two translation lists to evaluate the performance of the reranking module.

4.1 Baseline

We use the multi-word term translation technique of (Morin et al., 2007) as our baseline. A weakness of their approach (and most context-vector approaches) is that the translation lists produced do not have high precision, meaning that the correct translation is not often occupying the top position in the list. Conversely, a strength of that approach is that recall is high, meaning that most lists of reasonable length contain a correct translation.

Morin reports a precision of 60% for the top 20 candidates. We expect that our numbers using the same techniques would be lower due to the fact that the corpora in this experiment are fairly broad in topic, encompassing Natural Language Processing in general, while Morin's corpora had a very narrow domain.

4.1.1 Text Processing

To begin processing the Spanish corpus, we used ABBYY Finereader to OCR all the SEPLN papers. We then used the Spanish TreeTagger (Schmid, 1994) to lemmatize, POS-tag, and sentence-segment the Spanish text. For NP-chunking, we used a version of the MuNPEx chunker[3], modified to recognize more

[3]http://www.semanticsoftware.info/munpex

types of noun phrases. For the English corpus, we used the Stanford CoreNLP package (Toutanova et al., 2003) to segment, tokenize, and lemmatize the text.

4.1.2 Context Vectors

We consider the context of a word (or NP-chunk) to be all the words (or NP-chunks) appearing in the same sentence. This has been demonstrated in previous work to be a reasonable context. (Fung and Yee, 1998)

For each Spanish word appearing at least twice in the corpus, and for each of the English terms in the gold standard, we computed a context vector, representing the average context in which a word appears in its respective corpus.

We then used Mutual Information (Church and Hanks, 1989) to normalize these co-occurrences, in an attempt to keep the most frequent (and least significant) words from appearing in every context vector. We also considered TF-IDF and Log-Likelihood as potential normalization functions, but chose Mutual Information due to its superior performance. After normalization, we discarded all but the 250 most significant co-occurrences in each vector.

4.1.3 Translating the vectors

Using the Spanish-English dictionary described earlier, we translated each Spanish context vector into English. As in (Morin et al., 2007), when there were multiple translations listed for a word, we weighted the translations by their frequency in the English corpus.

4.1.4 Collecting Candidate Translations

For each English term, we selected the 250 most similar Spanish words (or NP-chunks) as measured by the cosine similarity between the English term's context vector and the translated Spanish context vector.

4.2 Reranking

In a domain with semi-structured and multilingual documents, we believe that the context-vector approach on its own misses out on a lot of potential translation clues. For example, if a Spanish document has an English title containing "Conditional Random Fields", a word in that paper is more likely to be a Spanish translation of that term than a randomly selected Spanish word. These two resources are much more likely to contain parallel words and phrases than would be expected at random.

We have extracted the following resources from the Spanish papers when available:

- English and Spanish titles
- English and Spanish abstracts
- Raw text
- English citation summaries
- Titles of citing English papers
- Titles of cited English papers
- English keywords

Table 2 shows the percentages of Spanish papers that had these features extracted.

Given a ranked list of translation candidates, we use the following method to rerank it, exploiting clues from the context in which the translation candidates appear.

For pairs of Spanish and English resources (R_{sp} and R_{en} respectively), we estimate the likelihood ratio of the probability of a word (or NP-chunk) w in R_{sp} being in the set of correct translations $tr(T)$ for term T

English title	41.4%
Spanish title	87.3%
English abstract	62.8%
Spanish abstract	77.6%
Citation summary	2.7%
Cited titles	60.2%
Citing titles	2.7%
English keywords	55.9%

Table 2: Percentage of the Spanish corpus that contained each extracted feature

in R_{en} versus the probability of a word randomly chosen from the lexicon L being correct by the following formula:

$$LogLikelihood_{R_{en},R_{sp}} = \\ log\left(\sum_{T\in\mathcal{T}} \frac{1}{|\mathcal{T}|} \frac{p(w \in tr(T)|w \in R_{sp}, T \in R_{en})}{p(w \in tr(T)|w \in L, T \in R_{en})}\right)$$

Intuitively, this measures how much more likely a word in the Spanish resource is to be a correct translation of a term in the English resource than a randomly selected word.

Table 3 shows the estimated log-likelihood ratios for pairs of English and Spanish resources. Certain pairs of passages, though in different languages, are still highly similar, and for the purposes of translation, are nearly parallel.

English resource	**Spanish resource**	**Log Likelihood**
Title	Title	6.90
	Abstract	4.94
	Raw text	2.76
Abstract	Title	5.94
	Abstract	4.89
	Raw text	2.38
Citation summary	Title	6.29
	Abstract	4.21
	Raw text	2.32
Cited title	Title	4.88
	Abstract	3.65
	Raw text	1.76
Citing title	Title	6.25
	Abstract	5.09
	Raw text	2.44
Keywords	Title	6.24
	Abstract	4.95
	Raw text	2.81

Table 3: Estimated log-likelihood ratios for pairs of Spanish and English resources. (These are estimated using all the available terms. The later evaluation splits the terms into two sets: one for parameter estimation and one for evaluation.)

To actually perform the reranking, we initially assign every item on the translation list a weight of zero. When a term satisfies the conditions of the previous list, the estimated log likelihood is added to its score. After this process, we arrive at a reranked list which we can compare with list produced by the context-vector approach alone. Since both lists contain exactly the same entries, we need only compare the orderings of the two lists.

5 Evaluation

For the English-Spanish language pair, we automatically extracted 1,080 pairs of English keywords and their Spanish translations from SEPLN papers. Many of the SEPLN papers listed identical keywords for the paper's topic in both languages. Figure 2 shows an example of keywords in Spanish and English. These keywords were also lemmatized in their respective languages to match the text in the corpora. We filtered out term translation pairs for which either the English term occurred fewer than 5 times or there was no Spanish translation which occurred at least twice. This left approximately 300 keywords with which to test.

Palabras clave: PLN, análisis automático, análisis profundo, gramática de análisis, representación semántica, catalán, español, inglés

Keywords: NLP, automatic parsing, deep parsing, parsing grammar, semantic representation, Catalan, Spanish, English

Figure 2: Examples of Spanish and English keywords in an SEPLN paper.

A translation candidate put forward by one of the systems was considered correct if it exactly matched one of the listed gold standard translations, notwithstanding inflectional variants or the inclusion/exclusion of a determiner.

For each of the baseline and reranking systems, we compiled the top 250 candidates produced by the system and compared the mean reciprocal rank (MRR) of each list.

To test, we used 10-fold cross validation, using 90% of the data each time to estimate the likelihood ratios and 10% of the data for calculating performance.

6 Results

For each system, we calculated the recall (the proportion of lists containing at least one correct translation) and the mean reciprocal rank (MRR), which measures how close to the top of the list the first correct translation occurs. As a clearer measure of the improvement yielded by reranking (since there can be no improvement on lists without a correct translation), we also calculated an adjusted MRR, which was the MRR calculated only over lists with a non-zero recall. These results are shown in Table 4.

System	Recall	MRR	adj. MRR
Context Vector	0.269	0.030	0.110
Reranking	0.269	0.172	0.640

Table 4: A comparison of the recall, MRR, and adjusted MRR for each system

For comparison to other papers, we also report top-N accuracy, that is, the proportion of lists that had a correct translation in the top-N. These numbers are presented in Table 5, both adjusted for zero-recall lists and unadjusted.

	Top 1	Top 5	Top 10	Top 20
Context Vector	0.010	0.038	0.071	0.112
Context Vector (adj)	0.035	0.141	0.259	0.417
Reranking	0.138	0.215	0.231	0.237
Reranking (adj)	0.512	0.798	0.857	0.881

Table 5: Top-N accuracy.

Figure 3 illustrates the results of the reranking process by showing the output of the context vector method and the output after reranking the list.

```
1   nondeterministic 0.064797185
2   that will 0.051721662
3   smt 0.05160014
4   clean 0.051406365
5   mayo 0.04894083
6   habitual 0.0454788
7   extra 0.045433186
8   tratar 0.045201465
9   otro lengua 0.04486432
10  vocabulario 0.044383436
11  tem 0.04420692
12  simulation 0.042538743
13  indicar 0.042119797
14  imp 0.040256664
15  un lengua 0.03919808
...
146 transductor de estado finitos 0.015205268
```

(a) Context vector candidate list

```
1  finitos 14.8054705103978
2  transductor 14.8054705103978
3  transductor de estado finitos 9.84106857789288
4  sordo 9.84106857789288
5  indicar 8.58141525183449
6  vocabulario 4.96440193250492
7  el vocabulario  4.96440193250492
8  conj  0
9  atts  0
10 vt  0
11 comprobación  0
12 lmt 0
13 desconocido 0
14 emborracha  0
15 mateos  0
16 el estudio  0
17 haver 0
18 un tipo 0
```

(b) Reranked candidate list

Figure 3: A comparison of the outputs of the two approaches. (a) shows the candidate translation list produced by the context vector method. The first correct translation appears at the 146th position in the list. (b) shows the reranked list. Here the first correct translation occurs at the 3rd position.

7 Discussion

7.1 Error Analysis

In analyzing the low recall of the context vector method, one issue concerned the normalization of co-occurrences. For many of the most commonly occurring English terms, these terms occurred so often that they co-occurred with many rare tokens. The normalization function did not seem to be able to determine significance well, and frequently selected a top-250 comprised entirely of improper English words, e.g., misspelled words, expressions from equations, or proper names. When computing cosine similarity to such a vector, every translated Spanish vector would have a similarity of 0, resulting in an empty translation candidate list. This issue occurred for 17% of the terms.

Another common issue was the appearance of English words in the list of candidate translations. In addition to having English titles or abstracts, some papers intermingled Spanish and English, and a few were even written primarily in English. Unfortunately, separating two intermingled languages can be a non-trivial problem. We estimate that English words comprised 22% of all the translation candidates.

As is evident in Figure 3, it was not uncommon for a constituent word in a multiword term to receive a higher score after reranking than the correct complete phrase. When the translation list did contain the correct multiword term, it ended up being ranked lower than one of its words 64% of the time. Just as constituent words were sometimes ranked too highly, commonly co-occurring words were sometimes erroneously promoted. Some English terms did not occur at all in any of the passages extracted for reranking; there were no clues available in these cases to rank one candidate above another. These three types of mistakes represent the majority of the errors made by the reranking method.

7.2 Context Vector Approach

Perhaps the biggest problem with the context vector method was simply the broadness of the domain. The context vector method has been proven quite effective in narrow domains, but in a domain as broad as NLP (though it's still very narrow compared to many potential domains!), there were simply too many similar contexts to choose from, especially for rarely occurring words that may not have appeared enough times to accurately represent their ideal context.

7.3 Reranking Approach

Though the reranking approach appears to be a substantial improvement over its baseline, it is worth noting that there is a potential bias in the results. The reranking method can only be tested for lists that contain at least one correct translation. It is possible that lists for which the context vector method produced a correct translation may not be representative of the list of terms as a whole. It may be that these are the "easier" cases, and the reranking approach could not be tested on the more difficult cases, though this is only speculation.

The results of a pilot experiment have also indicated that the reranking method is roughly as effective at translating Spanish terms into English, though we expect its context vector baseline to perform more poorly in this direction due to the larger English vocabulary size.

Referring back to Table 3, it is interesting to note that while the greatest likelihood ratio belongs to an English-title-to-Spanish-title alignment, the next greatest score actually comes from aligning the English citation summary with the Spanish title. So it seems that in the absence of truly parallel text, a citation summary can be nearly as salient as true translation of the title. The major drawback to this alignment however is the paucity of citations from English papers to Spanish papers. In fact, out of 1,135 SEPLN papers, only 31 were cited by any paper in the ACL Anthology.

It is also somewhat surprising that the titles of citing and cited papers are such strong sources of translations. To our knowledge, the textual similarity of titles between citing and cited papers has not been the subject of much research. These results indicate that such an effect may exist, and may in fact be preserved across language boundaries.

7.4 Future Directions

This approach may also have applications in domains other than scientific literature, specifically on the web. For example, a foreign language document on the web may be tweeted about in English or summarized in an English blog or even have useful English anchor text linking to it. Any of these resources could be used as nearly-parallel documents to provide clues for translations.

In future work, we would also like to test this approach with more languages, including languages like Chinese or Japanese that are very dissimilar to English. We would also like to incorporate more types of clues. Due to time constraints, we did not extract Spanish citing sentences and align them to English papers, but we believe there are a number potentially useful alignments between foreign language citing sentences and the papers that they cite. One that looks especially promising is co-citation.

Co-citation refers to the citation of the same paper by different authors, summarizing the paper's contributions in slightly different ways. Elkiss et al. showed that co-citation is highly correlated with textual

similarity (Elkiss et al., 2008). And in fact in preliminary experiments, we have found multilingual co-citation to be a very strong source of translation pairs.

The length of the translation candidate list necessary to achieve a decent recall and the success of the reranking method on a list of large size suggests that there is potential future work in adapting this approach to work without context vectors, i.e., the translation candidate list could simply be the entire lexicon. Such an approach would be advantageous because it would avoid the computational burden of calculating context vectors. It may also be beneficial if it can produce a higher recall on average than the context vector method.

8 Conclusion

Terminology can often be difficult to translate using machine translation methods that rely on parallel corpora, due to the fact that most terminology is only used in specific domains, which may not have a large amount of parallel text available. Using non-parallel scientific corpora in Spanish and English, we tested an algorithm for translation of terminology by comparing context vectors across languages.

We then proposed a method to improve the accuracy by reranking the candidates produced by that system by using clues from nearly-parallel passages from the corpora. This method greatly improved the accuracy and has shown that it is possible to exploit structure in multilingual scientific data to produce accurate translations of terminology even in the absence of large parallel corpora.

References

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2008)*.

Kenneth W. Church and Patrick Hanks. 1989. Word association norms, mutual information, and lexicography. In *ACM 27*.

Ido Dagan and Kenneth W. Church. 1994. Termight: Identifying and translating technical terminology. In *Proceedings of the 4th Conference on Applied Natural Language Processing*.

Aaron Elkiss, Siwei Shen, Anthony Fader, Güneş Erkan, David States, and Dragomir R Radev. 2008. Blind men and elephants: What do citation summaries tell us about the research article.

Pascale Fung and Kathleen R. McKeown. 1997. Finding terminology translations from non parallel corpora. In *Workshop On Very Large Corpora*.

Pascale Fung and Lo Yuen Yee. 1998. An ir approach for translating new words from nonparallel, comparable texts. In *COLING ACL*.

Dekang Lin, Shaojun Zhao, Benjamin van Durme, and Marius Pasca. 2008. Mining parenthetical translations from the web by word alignment. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*.

Adam Lopez. 2008. Statistical machine translation. *ACM Computing Surveys*, 40(3).

Emmanuel Morin, Béatrice Daille, Koichi Takeuchi, and Kyo Kageura. 2007. Bilingual terminology mining - using brain, not brawn comparable corpora. In *45th Annual Meeting of the Association of Computational Linguistics*.

Drafos Stefan Munteanu and Daniel Marcu. 2006. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–505.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

Frank Smadja, Kathleen R McKeown, and Vasileios Hatzsivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 24(4):1–38.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL*, pages 252–259.

Francis M Tyers, Felipe Sánchez-Martínez, Sergio Ortiz-Rojas, and Mikel L Forcada. 2010. Free/open-source resources in the apertium platform for machine translation research and development.

Xianchao Wu, Naoaki Okazaki, and Jun'ichi Tsujii. 2004. Semi-supervised lexicon mining from parenthetical expressions in monolingual web pages. In *HLT NAACL*.