Word-for-Word Glossing with Contextually Similar Words

Patrick Pantel and Dekang Lin
Department of Computer Science
University of Manitoba
Winnipeg, Manitoba R3T 2N2 Canada
{ppantel, lindek}@cs.umanitoba.ca

Abstract

Many corpus-based machine translation systems require parallel corpora. In this paper, we present a word-for-word glossing algorithm that requires only a source language corpus. To gloss a word, we first identify its similar words that occurred in the same context in a large corpus. We then determine the gloss by maximizing the similarity between the set of contextually similar words and the different translations of the word in a bilingual thesaurus.

1. Introduction

Word-for-word glossing is the process of directly translating each word or term in a document without considering the word order. Automating this process would benefit many NLP applications. For example, in cross-language information retrieval, glossing a document often provides a sufficient translation for humans to comprehend the key concepts. Furthermore, a glossing algorithm can be used for lexical selection in a full-fledged machine translation (MT) system.

Many corpus-based MT systems require parallel corpora (Brown et al., 1990; Brown et al., 1991; Gale and Church, 1991; Resnik, 1999). Kikui (1999) used a word sense disambiguation algorithm and a non-parallel bilingual corpus to resolve translation ambiguity.

In this paper, we present a word-for-word glossing algorithm that requires only a source language corpus. The intuitive idea behind our algorithm is the following. Suppose w is a word to be translated. We first identify a set of words similar to w that occurred in the same context as w in a large corpus. We then use this set (called the **contextually similar words** of w) to select a translation for w. For example, the contextually similar words of *duty* in *fiduciary duty* include *responsibility*, *obligation*, *role*, ... This list is then used to select a translation for *duty*.

In the next section, we describe the resources required by our algorithm. In Section 3, we present an algorithm for constructing the contextually similar words of a word in a context. Section 4 presents the word-for-word glossing algorithm and Section 5 describes the group similarity metric used in our algorithm. In Section 6, we present some experimental results and finally, in Section 7, we conclude with a discussion of future work.

2. Resources

The input to our algorithm includes a collocation database (Lin, 1998b) and a corpus-based thesaurus (Lin, 1998a), which are both available on the Internet[1]. In addition, we require a bilingual thesaurus. Below, we briefly describe these resources.

2.1. Collocation database

Given a word w in a dependency relationship (such as *subject* or *object*), the collocation database can be used to retrieve the words that occurred in that relationship with w, in a large corpus, along with their frequencies[2]. Figure 1 shows excerpts of the entries in the collocation database for the words *corporate*, *duty*, and *fiduciary*. The database contains a total of 11 million unique dependency relationships.

[1] Available at www.cs.umanitoba.ca/~lindek/depdb.htm and www.cs.umanitoba.ca/~lindek/simdb.htm

[2] We use the term *collocation* to refer to a pair of words that occur in a dependency relationship (rather than the linear proximity of a pair of words).

Table 1. Clustered similar words of *duty* as given by (Lin, 1998a).

CLUSTER	CLUSTERED SIMILAR WORDS OF *DUTY* (WITH SIMILARITY SCORE)
1	responsibility 0.16, obligation 0.109, task 0.101, function 0.098, role 0.091, post 0.087, position 0.086, job 0.084, chore 0.08, mission 0.08, assignment 0.079, liability 0.077, ...
2	tariff 0.091, restriction 0.089, tax 0.086, regulation 0.085, requirement 0.081, procedure 0.079, penalty 0.079, quota 0.074, rule 0.07, levy 0.061, ...
3	fee 0.085, salary 0.081, pay 0.064, fine 0.058
4	personnel 0.073, staff 0.073
5	training 0.072, work 0.064, exercise 0.061
6	privilege 0.069, right 0.057, license 0.056

2.2. Corpus-based thesaurus

Using the collocation database, Lin used an unsupervised method to construct a corpus-based thesaurus (Lin, 1998a) consisting of 11839 nouns, 3639 verbs and 5658 adjectives/adverbs. Given a word *w*, the thesaurus returns a clustered list of similar words of *w* along with their similarity to *w*. For example, the clustered similar words of *duty* are shown in Table 1.

2.3. Bilingual thesaurus

Using the corpus-based thesaurus and a bilingual dictionary, we manually constructed a bilingual thesaurus. The entry for a source language word *w* is constructed by manually associating one or more clusters of similar words of *w* to each candidate translation of *w*. We refer to the assigned clusters as **Words Associated with a Translation (WAT)**. For example, Figure 2 shows an excerpt of our *English/French* bilingual thesaurus for the words *account* and *duty*.

Although the WAT assignment is a manual process, it is a considerably easier task than providing lexicographic definitions. Also, we only require entries for source language words that have multiple translations. In Section 7, we discuss a method for automatically assigning the WATs.

corporate:

modifier-of:	client 196, debt 236, development 179, fee 6, function 16, headquarter 316, IOU 128, levy 3, liability 14, manager 203, market 195, obligation 1, personnel 7, profit 595, responsibility 27, rule 7, staff 113, tax 201, training 2, vice president 231, ...

duty:

object-of:	assume 177, breach 111, carry out 71, do 114, have 257, impose 114, perform 151, ...
subject-of:	affect 4, apply 6, include 42, involve 8, keep 5, officer 22, protect 8, require 13, ...
adj-modifier:	active 202, additional 46, administrative 44, fiduciary 317, official 66, other 83, ...

fiduciary:

modifier-of:	act 2, behavior 1, breach 2, claim 1, company 2, duty 317, irresponsibility 2, obligation 32, requirement 1, responsibility 89, role 2, ...

Figure 1. Excepts of entries in the collocation database for the words *corporate*, *duty*, and *fiduciary*.

account:

1. compte:	fund, deposit, loan, asset, portfolio, investment, transaction, payment, saving, money, contract, Budget, reserve, security, contribution, debt, property, holding, interest, bond, plan, business, ...
2. rapport:	report, statement, testimony, card, story, record, document, data, information, view, check, figure, article, description, estimate, assessment, number, statistic, comment, letter, picture, note, ...

duty:

1. devoir:	responsibility, obligation, task, function, role, post, position, job, chore, mission, assignment, liability, ...
2. taxe:	tariff, restriction, tax, regulation, requirement, procedure, penalty, quota, rule, levy, ...

WAT for *compte*

Figure 2. Bilingual thesaurus entries for *account* and *duty*.

3. Contextually Similar Words

The contextually similar words of a word *w* are words similar to the intended meaning of *w* in its context. Figure 3 gives the data flow diagram for our algorithm for identifying the contextually similar words of *w*. Data are represented by ovals, external resources by double ovals and processes by rectangles.

By parsing a sentence with *Minipar*[3], we extract the dependency relationships involving *w*. For each dependency relationship, we retrieve

[3] Available at www.cs.umanitoba.ca/~lindek/minipar.htm

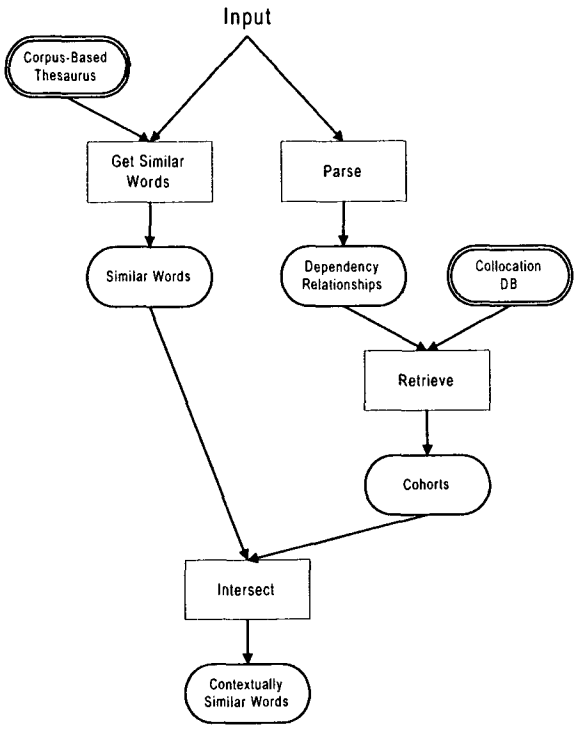


Figure 3. Data flow diagram for identifying the contextually similar words of a word in context.

from the collocation database the words that occurred in the same dependency relationship as *w*. We refer to this set of words as the **cohort** of *w* for that dependency relationship. Consider the word *duty* in the contexts *corporate duty* and *fiduciary duty*. The cohort of *duty* in *corporate duty* consists of nouns modified by *corporate* in Figure 1 (e.g. client, debt, development, ...) and the cohort of *duty* in *fiduciary duty* consists of nouns modified by *fiduciary* in Figure 1 (e.g. act, behaviour, breach, ...).

Intersecting the set of similar words and the cohort then forms the set of contextually similar words of *w*. For example, Table 2 shows the contextually similar words of *duty* in the contexts *corporate duty* and *fiduciary duty*. The words in the first row are retrieved by intersecting the words in Table 1 with the nouns modified by *corporate* in Figure 1. Similarly, the second row represents the intersection of the words in Table 1 and the nouns modified by *fiduciary* in Figure 1.

The first set of contextually similar words in Table 2 contains words that are similar to both

Table 2. The words similar to *duty* that occurred in the contexts *corporate duty* and *fiduciary duty*.

CONTEXT	CONTEXTUALLY SIMILAR WORDS OF *DUTY*
corporate duty	fee, function, levy, liability, obligation, personnel, responsibility, rule, staff, tax, training
fiduciary duty	obligation, requirement, responsibility, role

the *responsibility* and *tax* senses of *duty*, reflecting the fact that the meaning of *duty* is indeed ambiguous if *corporate duty* is its sole context. In contrast, the second row in Table 2 clearly indicates the responsibility sense of *duty*.

While previous word sense disambiguation algorithms rely on a lexicon to provide sense inventories of words, the contextually similar words provide a way of distinguishing between different senses of words without committing to any particular sense inventory.

4. Overview of the Word-for-Word Glossing Algorithm

Figure 4 illustrates the data flow of the word-for-word glossing algorithm and Figure 5 describes it.

For example, suppose we wish to translate into French the word *duty* in the context *corporate fiduciary duty*. **Step 1** retrieves the candidate translations for *duty* and its WATs from Figure 2. In **Step 2**, we construct two lists of contextually similar words, one for the dependency context *corporate duty* and one for the dependency context *fiduciary duty*, shown in Table 2. The proposed translation for the context is obtained by maximizing the group similarities between the lists of contextually similar words and the WATs.

Using the group similarity measure from Section 5, Table 3 lists the group similarity scores between each list of contextually similar words and each WAT as well as the final combined score for each candidate translation. The combined score for a candidate is the sum of the logs of all group similarity scores involving its WAT. The correct proposed translation for *duty* in this context is *devoir* since its WAT received the highest score.

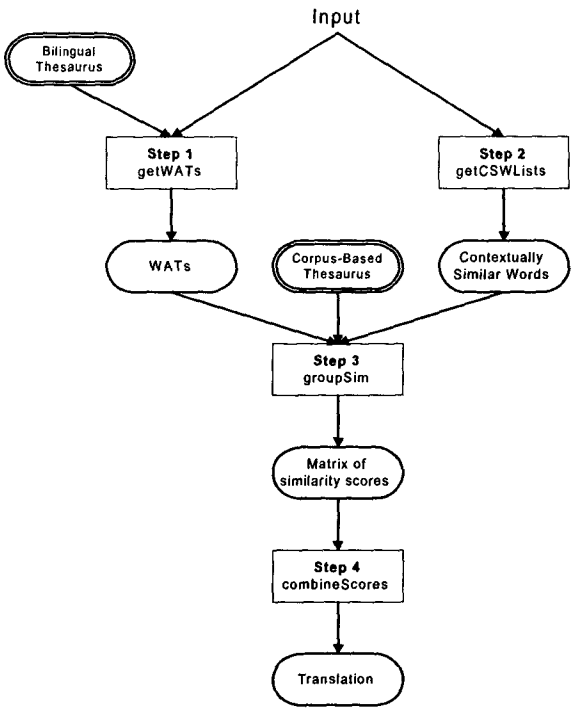


Figure 4. Data flow diagram for the word-for-word glossing algorithm.

Input:	A word w to be translated and a set of dependency contexts involving w.
Step 1:	Retrieve the candidate translations of w and the corresponding WATs from the bilingual thesaurus.
Step 2:	Find the contextually similar words of w in each dependency context using the algorithm from Section 3.
Step 3:	Compute the group similarity (see details in Section 5) between each set of contextually similar words and each WAT; the results are stored in a matrix t, where $t[i, j]$ is the group similarity between the i^{th} list of contextually similar words and the j^{th} WAT.
Step 4:	Add the logs of the group similarity scores in column of t to obtain a score for each WAT.
Output:	The candidate translation corresponding to the WAT with the highest score.

Figure 5. The word-for-word glossing algorithm.

5. Group Similarity

The corpus-based thesaurus contains only the similarities between individual pairs of words. In our algorithm, we require the similarity between groups of words. The group similarity measure we use is proposed by Karypis et al. (1999). It takes as input two groups of elements, G_1 and G_2, and a similarity matrix, *sim*, which specifies the similarity between individual elements. G_1 and G_2 are describable by graphs where the vertices are the words and each weighted edge between vertices w_1 and w_2 represents the similarity, $sim(w_1, w_2)$, between the words w_1 and w_2.

Karypis et al. consider both the interconnectivity and the closeness of the groups. The **absolute interconnectivity** between G_1 and G_2, $AI(G_1, G_2)$, is defined as the aggregate similarity between the two groups:

$$AI(G_1, G_2) = \sum_{x \in G_1} \sum_{y \in G_2} sim(x, y)$$

The **absolute closeness** between G_1 and G_2, $AC(G_1, G_2)$, is defined as the average similarity between a pair of elements, one from each group:

$$AC(G_1, G_2) = \frac{1}{|G_1||G_2|} AI(G_1, G_2)$$

Table 3. Group similarity scores between the contextually similar words of *duty* in *corporate duty* and *fiduciary duty* with the WATs for candidate translations *devoir* and *taxe*.

	CANDIDATE *DEVOIR*	CANDIDATE *TAXE*
corporate duty	60.3704	16.569
fiduciary duty	51.2960	4.8325
Combined Score	8.0381	4.3829

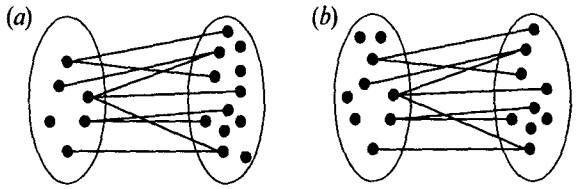


Figure 6. An example illustrating the difference between the interconnectivity and closeness measures. The interconnectivity in (*a*) and (*b*) remains constant while the closeness in (*a*) is higher than in (*b*) since there are more zero similarity pairs in (*b*).

Table 4. Candidate translations for each testing word along with their frequency of occurrence in the test corpus.

WORD	CANDIDATE TRANSLATION	ENGLISH SENSE	FREQUENCY OF OCCURRENCE
account	compte	bank account, business	245
	rapport	report, statement	55
duty	devoir	responsibility, obligation	80
	taxe	tax	30
race	course	contest	87
	race	racial group	23
suit	procès	lawsuit	281
	costume	garment	17
check	chèque	draft, bank order	105
	contrôle	evaluation, verification	25
record	record	unsurpassed statistic/performance	98
	enregistrement	recorded data or documentation	12

The difference between the absolute interconnectivity and the absolute closeness is that the latter takes zero similarity pairs into account. In Figure 6, the interconnectivity in (*a*) and (*b*) remains constant. However, the closeness in (*a*) is higher than in (*b*) since there are more zero similarity pairs in (*b*).

Karypis et al. normalized the absolute interconnectivity and closeness by the internal interconnectivity and closeness of the individual groups. The normalized measures are referred to as **relative interconnectivity**, $RI(G_1, G_2)$, and **relative closeness**, $RC(G_1, G_2)$. The internal interconnectivity and closeness are obtained by first computing a **minimal edge bisection** of each group. An even-sized partition $\{G', G''\}$ of a group G is called a minimal edge bisection of G if $AI(G', G'')$ is minimal among all such partitions. The **internal interconnectivity** of G, $II(G)$, is defined as $II(G) = AI(G', G'')$ and the **internal closeness** of G, $IC(G)$, as $IC(G) = AC(G', G'')$.

Minimal edge bisection is performed for all WATs and all sets of contextually similar words. However, the minimal edge bisection problem is *NP*-complete (Garey and Johnson, 1979). Fortunately, state of the art graph partitioning algorithms can approximate these bisections in polynomial time (Goehring and Saad, 1994; Karypis and Kumar, 1999; Kernighan and Lin, 1970). We used the same approximation methods as in (Karypis et al., 1999).

The similarity between G_1 and G_2 is then defined as follows:

$$groupSim(G_1, G_2) = RI(G_1, G_2) \times RC(G_1, G_2)$$

where

$$RI(G_1, G_2) = \frac{2AI(G_1, G_2)}{II(G_1) + II(G_2)}$$

is the relative interconnectivity and

$$RC(G_1, G_2) = \frac{AC(G_1, G_2)}{\frac{|G_1|}{|G_1|+|G_2|} IC(G_1) + \frac{|G_2|}{|G_1|+|G_2|} IC(G_2)}$$

is the relative closeness.

6. Experimental Results

The design of our glossing algorithm is applicable to any source/destination language pair as long as a source language parser is available. We considered *English-to-French* translations in our experiments.

We experimented with six English nouns that have multiple French translations: *account*, *duty*, *race*, *suit*, *check*, and *record*. Using the 1987 Wall Street Journal files on the LDC/DCI CD-

ROM, we extracted a testing corpus[4] consisting of the first 100 to 300 sentences containing the non-idiomatic usage of the six nouns[5]. Then, we manually tagged each sentence with one of the candidate translations shown in Table 4.

Each noun in Table 4 translates more frequently to one candidate translation than the other. In fact, always choosing the candidate *procès* as the translation for *suit* yields 94% accuracy. A better measure for evaluating the system's classifications considers both the algorithm's precision and recall on each candidate translation. Table 5 illustrates the precision and recall of our glossing algorithm for each candidate translation. Albeit precision and recall are used to evaluate the quality of the classifications, overall accuracy is sufficient for comparing different approaches with our system.

In Section 3, we presented an algorithm for identifying the contextually similar words of a word in a context using a corpus-based thesaurus and a collocation database. Each of the six nouns has similar words in the corpus-based thesaurus. However, in order to find contextually similar words, at least one similar word for each noun must occur in the collocation database in a given context. Thus, the algorithm for constructing contextually similar words is dependent on the coverage of the collocation database. We estimated this coverage by counting the number of times each of the six nouns, in several different contexts, has at least one contextually similar word. The result is shown in Table 6.

In Section 5, we described a group similarity metric, *groupSim*, which we use for comparing a WAT with a set of contextually similar words. In Figure 7, we compare the translation accuracy of our algorithm using other group similarity metrics. Suppose G_1 and G_2 are two groups of words and w is the word that we wish to translate. The metrics used are:

1. ***closest3:***
   sum of similarity of the three closest pairs of words from each group.

Table 5. Precision vs. Recall for each candidate translation.

WORD	CANDIDATE	PRECISION	RECALL
account	compte	0.982	0.902
	rapport	0.680	0.927
duty	devoir	0.951	0.963
	taxe	0.897	0.867
race	course	0.945	0.989
	race	0.947	0.783
suit	procès	0.996	0.993
	costume	0.889	0.941
check	chèque	0.951	0.924
	contrôle	0.714	0.800
record	record	0.968	0.918
	enregistrement	0.529	0.750

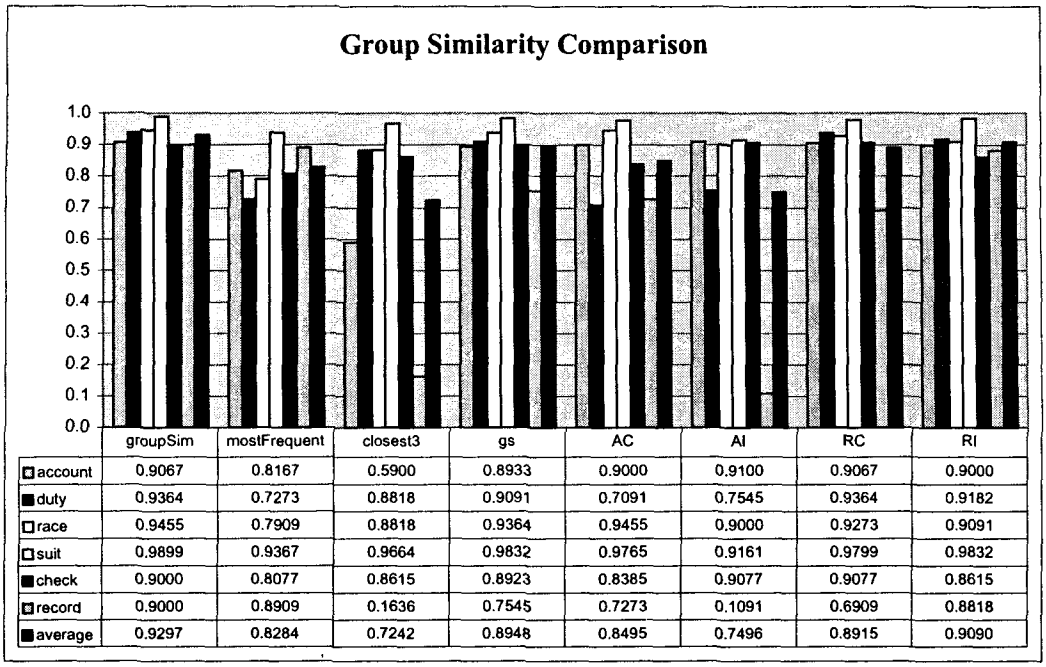
Table 6. The coverage of the collocation database, shown by the frequency with which a word in a given context has at least one contextually similar word.

WORD	NUMBER OF CONTEXTS	COVERAGE
account	1074	95.7%
duty	343	93.3%
race	294	92.5%
suit	332	91.9%
check	2519	87.5%
record	1655	92.8%

2. ***gs***:
$$\frac{\sum_{x\in G_1} sim(x,w)\times \max_{y\in G_2} sim(x,y) + \sum_{y\in G_2} sim(y,w)\times \max_{x\in G_1} sim(y,x)}{\sum_{x\in G_1} sim(x,w) + \sum_{y\in G_2} sim(y,w)}$$
3. ***AC***:
   as defined in Section 5.
4. ***AI***:
   as defined in Section 5.
5. ***RC***:
   as defined in Section 5.
6. ***RI***:
   as defined in Section 5.

[4] Available at ftp.cs.umanitoba.ca/pub/ppantel/download/wfwgtest.zip

[5] Omitted idiomatic phrases include *take into account, keep in check, check out, ...*



	groupSim	mostFrequent	closest3	gs	AC	AI	RC	RI
account	0.9067	0.8167	0.5900	0.8933	0.9000	0.9100	0.9067	0.9000
duty	0.9364	0.7273	0.8818	0.9091	0.7091	0.7545	0.9364	0.9182
race	0.9455	0.7909	0.8818	0.9364	0.9455	0.9000	0.9273	0.9091
suit	0.9899	0.9367	0.9664	0.9832	0.9765	0.9161	0.9799	0.9832
check	0.9000	0.8077	0.8615	0.8923	0.8385	0.9077	0.9077	0.8615
record	0.9000	0.8909	0.1636	0.7545	0.7273	0.1091	0.6909	0.8818
average	0.9297	0.8284	0.7242	0.8948	0.8495	0.7496	0.8915	0.9090

Figure 7. Performance comparison of different group similarity metrics.

In *mostFrequent*, we include the results obtained if we always choose the translation that occurs most frequently in the testing corpus.

We also compared the accuracy of our glossing algorithm with *Systran*'s translation system by feeding the testing sentences into Systran's web interface[6] and manually examining the results. Figure 8 summarizes the overall accuracy obtained by each system and the baseline on the testing corpus. Systran tended to prefer one candidate translation over the other and committed the majority of its errors on the non-preferred senses. Consequently, Systran is very accurate if its preferred sense is the frequent sense (as in *account* and *duty*) but is very inaccurate if its preferred sense is the infrequent one (as in *race*, *suit*, and *check*).

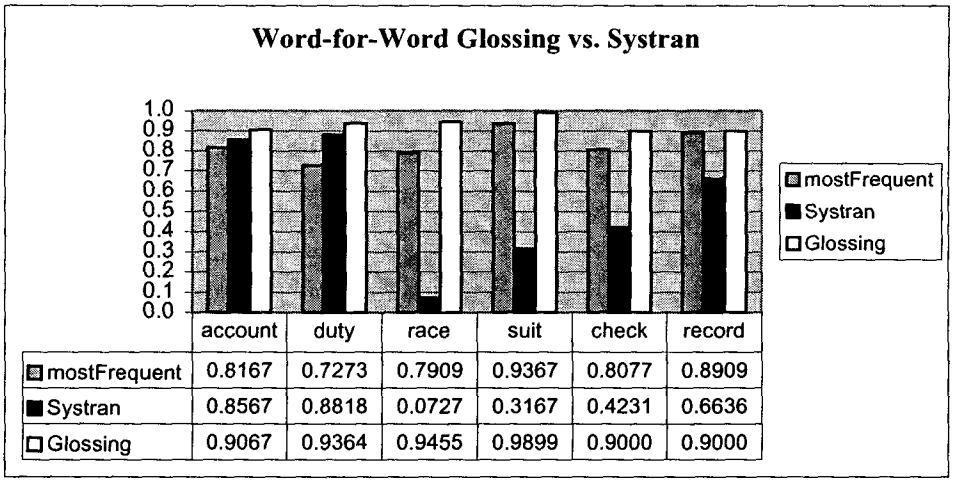
7. Conclusion and Future Work

This paper presents a word-for-word glossing algorithm. The gloss of a word is determined by maximizing the similarity between the set of contextually similar words and the different translations of the word in a bilingual thesaurus.

[6] Available at babelfish.altavista.com/cgi-bin/translate

The algorithm presented in this paper can be improved and extended in many ways. At present, our glossing algorithm does not take the prior probabilities of translations into account. For example, in WSJ, the bank account sense of *account* is much more common than the report sense. We should thus tend to prefer this sense of *account*. This is achievable by weighting the translation scores by the prior probabilities of the translations. We are investigating an *Expectation-Maximization* (EM) (Dempster et al., 1977) algorithm to learn these prior probabilities. Initially, we assume that the candidate translations for a word are uniformly distributed. After glossing each word in a large corpus, we refine the prior probabilities using the frequency counts obtained. This process is repeated several times until the empirical prior probabilities closely approximate the true prior probabilities.

Finally, as discussed in Section 2.3, automatically constructing the bilingual thesaurus is necessary to gloss whole documents. This is attainable by adding a corpus-based destination language thesaurus to our system. The process of assigning a cluster of similar words as a WAT to a candidate translation c is as follows. First, we



	account	duty	race	suit	check	record
mostFrequent	0.8167	0.7273	0.7909	0.9367	0.8077	0.8909
Systran	0.8567	0.8818	0.0727	0.3167	0.4231	0.6636
Glossing	0.9067	0.9364	0.9455	0.9899	0.9000	0.9000

Figure 8. Performance comparison of the word-for-word glossing algorithm and Systran.

automatically obtain the candidate translations for a word using a bilingual dictionary. With the destination language thesaurus, we obtain a list S of all words similar to c. With the bilingual dictionary, replace each word in S by its source language translations. Using the group similarity metric from Section 5, assign as the WAT the cluster of similar words (obtained from the source language thesaurus) most similar to S.

Acknowledgements

The authors wish to thank the reviewers for their helpful comments. This research was partly supported by Natural Sciences and Engineering Research Council of Canada grants OGP121338 and PGSA207797.

References

Peter F. Brown; John Cocke; Stephen A. Della Pietra; Vincent J. Della Pietra; Fredrick Jelinek; John D. Lafferty; Robert L. Mercer and Paul S. Roossin. 1990. *A Statistical Approach to Machine Translation.* Computation Linguistics, 16(2).

Peter F. Brown; Jennifer C. Lai and Robert L. Mercer. 1991. *Aligning Sentences in Parallel Corpora.* In Proceedings of ACL91. Berkeley.

A. P. Dempster; N. M. Laird; & D. B. Rubin. 1977. *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society, Series B, 39(1).

W. A. Gale and K. W. Church. 1991. *A Program for Aligning Sentences in Bilingual Corpora.* In Proceedings of ACL91. Berkeley.

M. R. Garey and D. S. Johnson. 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W. H. Freeman.

T. Goehring and Y. Saad. 1994. *Heuristic Algorithms for Automatic Graph Partitioning.* Technical Report. Department of Computer Science, University of Minnesota.

George Karypis and Vipin Kumar. 1999. *A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs.* SIAM Journal on Scientific Computing, 20(1).

George Karypis; Eui-Hong Han and Vipin Kumar. 1999. *Chameleon: A Hierarchical Clustering Algorithm Using Dynamic Modeling.* IEEE Computer: Special Issue on Data Analysis and Mining, 32(8). http://www-users.cs.umn.edu/~karypis/publications/Papers/PDF/chameleon.pdf

B. W. Kernighan and S. Lin. 1970. *An Efficient Heuristic Procedure for Partitioning Graphs.* The Bell System Technical Journal.

Genichiro Kikui. 1999. *Resolving Translation ambiguity using Non-parallel Bilingual Corpora.* In Proceedings of ACL99 Workshop on Unsupervised Learning in Natural Language Processing.

Dekang Lin. 1998a. *Automatic Retrieval and Clustering of Similar Words.* In Proceedings of COLING-ACL98. Montreal, Canada.

Dekang Lin. 1998b. *Extracting Collocations from Text Corpora.* Workshop on Computational Terminology. Montreal, Canada.

Philip Resnik. 1999. *Mining the Web for Bilingual Text.* In Proceedings of ACL99. College Park, Maryland.