WORD-SENSE DISAMBIGUATION USING STATISTICAL METHODS

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer

IBM Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

ABSTRACT

We describe a statistical technique for assigning senses to words. An instance of a word is assigned a sense by asking a question about the context in which the word appears. The question is constructed to have high mutual information with the translation of that instance in another language. When we incorporated this method of assigning senses into our statistical machine translation system, the error rate of the system decreased by thirteen percent.

INTRODUCTION

An alluring aspect of the statistical approach to machine translation rejuvenated by Brown *et al.* [Brown *et al.*, 1988, Brown *et al.*, 1990] is the systematic framework it provides for attacking the problem of lexical disambiguation. For example, the system they describe translates the French sentence *Je vais prendre la décision* as *I will make the decision*, correctly interpreting *prendre* as *make*. The statistical translation model, which supplies English translations of French words, prefers the more common translation *take*, but the trigram language model recognizes that the three-word sequence *make the decision* is much more probable than *take the decision*.

The system is not always so successful. It incorrectly renders *Je vais prendre ma propre décision* as *I will take my own decision*. The language model does not realize that *take my own decision* is improbable because *take* and *decision* no longer fall within a single trigram.

Errors such as this are common because the statistical models only capture local phenomena; if the context necessary to determine a translation falls outside the scope of the models, the word is likely to be translated incorrectly. However, if the relevant context is encoded locally, the word should be translated correctly. We can achieve this within the traditional paradigm of analysis, transfer, and synthesis by incorporating into the analysis phase a sense-disambiguation component that assigns sense labels to French words. If *prendre* is labeled with one sense in the context of *décision* but with a different sense in other contexts, then the translation model will learn from training data that the first sense usually translates to *make*, whereas the other sense usually translates to *take*.

Previous efforts at algorithmic disambiguation of word senses [Lesk, 1986, White, 1988, Ide and Véronis, 1990] have concentrated on information that can be extracted from electronic dictionaries, and focus, therefore, on senses as determined by those dictionaries. Here, in contrast, we present a procedure for constructing a sense-disambiguation component that labels words so as to elucidate their translations in another language. We are con-

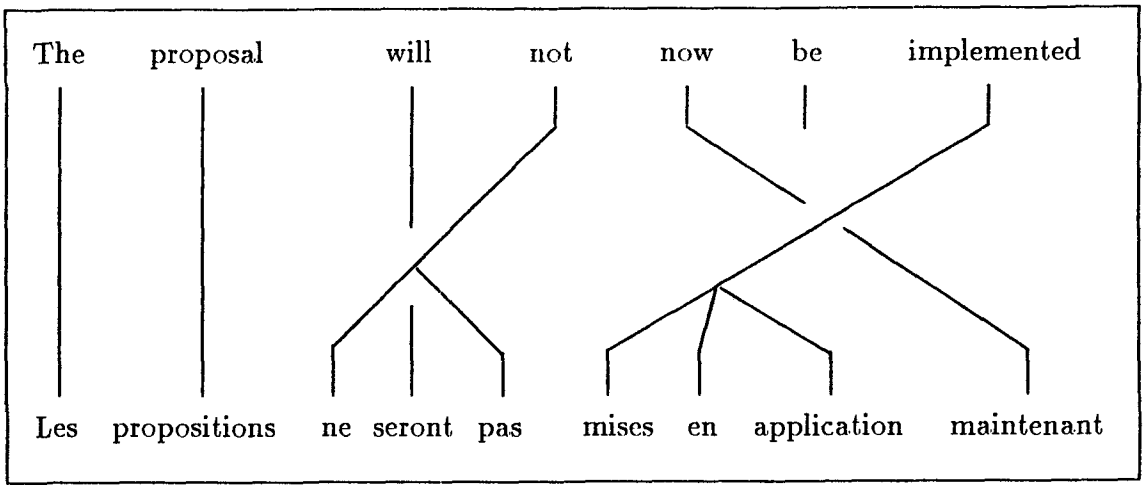

Figure 1: Alignment Example

cerned about senses as they occur in a dictionary only to the extent that those senses are translated differently. The French noun *intérêt*, for example, is translated into German as either *Zins* or *Interesse* according to its sense, but both of these senses are translated into English as *interest*, and so we make no attempt to distinguish them.

STATISTICAL TRANSLATION

Following Brown *et al.* [Brown *et al.*, 1990], we choose as the translation of a French sentence F that sentence E for which $\Pr(E|F)$ is greatest. By Bayes' rule,

$$\Pr(E|F) = \frac{\Pr(E)\Pr(F|E)}{\Pr(F)}. \qquad (1)$$

Since the denominator does not depend on E, the sentence for which $\Pr(E|F)$ is greatest is also the sentence for which the product $\Pr(E)\Pr(F|E)$ is greatest. The first factor in this product is a statistical characterization of the English language and the second factor is a statistical characterization of the process by which English sentences are translated into French. We can compute neither factors precisely. Rather, in statistical translation, we employ models from which we can obtain estimates of these values. We call the model from which we compute $\Pr(E)$ the language model and that from which we compute $\Pr(F|E)$ the translation model.

The translation model used by Brown *et al.* [Brown *et al.*, 1990] incorporates the concept of an *alignment* in which each word in E acts independently to produce some of the words in F. If we denote a typical alignment by A, then we can write the probability of F given E as a sum over all possible alignments:

$$\Pr(F|E) = \sum_A \Pr(F, A|E). \qquad (2)$$

Although the number of possible alignments is a very rapidly growing function of the lengths of the French and English sentences, only a tiny fraction of the alignments contributes substantially to the sum, and of these few, one makes the greatest contribution. We call this most probable alignment the *Viterbi alignment* between E and F.

The identity of the Viterbi alignment for a pair of sentences depends on the details of the translation model, but once the model is known, probable alignments can be discovered algorithmically [Brown *et al.*, 1991]. Brown *et al.* [Brown *et al.*, 1990], show an example of such an automatically derived alignment in their Figure 3. (For the reader's convenience, we have reproduced that figure here as Figure 1.)

In a Viterbi alignment, a French word that is connected by a line to an English word is said to be *aligned* with that English word. Thus, in Figure 1, *Les* is aligned with *The*, *propositions* with *proposal*, and so on. We call a pair of aligned words obtained in this way a *connection*.

From the Viterbi alignments for 1,002,165 pairs of short French and English sentences from the Canadian Hansard data [Brown *et al.*, 1990], we have extracted a set of 12,028,485 connections. Let $p(e, f)$ be the probability that a connection chosen at random from this set will connect the English word e to the French word f. Because each French word gives rise to exactly one connection, the right marginal of this distribution is identical to the distribution of French words in these sentences. The left marginal, however, is not the same as the distribution of English words: English words that tend to produce several French words at a time are overrepresented while those that tend to produce no French words are underrepresented.

SENSES BASED ON BINARY QUESTIONS

Using $p(e, f)$ we can compute the mutual information between a French word and its English mate in a connection. In this section, we discuss a method for labelling a word with a sense that depends on the context in which it appears in such a way as to increase the mutual information between the members of a connection.

In the sentence *Je vais prendre ma propre décision*, the French verb *prendre* should be translated as *make* because the object of *prendre* is *décision*. If we replace *décision* by *voiture*, then *prendre* should be translated as *take* to yield *I will take my own car*. In these examples, one can imagine assigning a sense to *prendre* by asking whether the first noun to the right of *prendre* is *décision* or *voiture*. We say that the noun to the right is the *informant* for *prendre*.

In *Il doute que les nôtres gagnent*, which means *He doubts that we will win*, the French word *il* should be translated as *he*. On the other hand, in *Il faut que les nôtres gagnent*, which means *It is necessary that we win*, *il* should be translated as *it*. Here, we can determine which sense to assign to *il* by asking about the identity of the first verb to its right. Even though we cannot hope to determine the translation of *il* from this informant unambiguously, we can hope to obtain a significant amount of information about the translation.

As a final example, consider the English word *is*. In the sentence *I think it is a problem*, it is best to translate *is* as *est* as in *Je pense que c'est un problème*. However, this is certainly not true in the sentence *I think there is a problem*, which translates as *Je pense qu'il y a un problème*. Here we can reduce the entropy of the distribution of the translation of *is* by asking if the word to the left is *there*. If so, then *is* is less likely to be translated as *est* than if not.

Motivated by examples like these, we investigated a simple method of assigning two senses to a word w by asking a single binary question about one word of the context in which w appears. One does not know beforehand whether the informant will be the first noun to the right, the first verb to the right, or some other word in the context of w. However, one can construct a question for each of a number of candidate informant sites, and then choose the most informative question.

Given a potential informant such as the first noun to the right, we can construct a question that has high mutual information with the translation of w by using the *flip-flop* algorithm devised by Nadas, Nahamoo, Picheny, and Powell [Nadas *et al.*, 1991]. To understand their algorithm, first imagine that w is a French word and that English words which are possible translations of w have been divided into two classes. Consider the problem of constructing a binary question about the potential informant that provides maximal information about these two English word classes. If the French vocabulary is of size V, then there

are 2^V possible questions. However, using the splitting theorem of Breiman, Friedman, Olshen, and Stone [Breiman *et al.*, 1984], it is possible to find the most informative of these 2^V questions in time which is linear in V.

The flip-flop algorithm begins by making an initial assignment of the English translations into two classes, and then uses the splitting theorem to find the best question about the potential informant. This question divides the French vocabulary into two sets. One can then use the splitting theorem to find a division of the English translations of w into two sets which has maximal mutual information with the French sets. In the flip-flop algorithm, one alternates between splitting the French vocabulary into two sets and the English translations of w into two sets. After each such split, the mutual information between the French and English sets is at least as great as before the split. Since the mutual information is bounded by one bit, the process converges to a partition of the French vocabulary that has high mutual information with the translation of w.

A PILOT EXPERIMENT

We used the flip-flop algorithm in a pilot experiment in which we assigned two senses to each of the 500 most common English words and two senses to each of the 200 most common French words.

For a French word, we considered questions about seven informants: the word to the left, the word to the right, the first noun to the left, the first noun to the right, the first verb to the left, the first verb to the right, and the tense of either the current word, if it is a verb, or of the first verb to the left of the current word. For an English word, we only considered questions about the the word to the left and the word two to the left. We restricted the English questions to the previous two words so that we could easily use them in our translation system which produces an English sentence from left to right. When a potential informant did not exist, because, say there was no noun to the left of some word in a particular sentence, we used the special word, *TERM_WORD*. To find the nouns and verbs in our French sentences, we used the tagging algorithm described by Merialdo [Merialdo, 1990].

Word:	prendre
Informant:	Right noun
Information:	.381 bits

Sense 1	Sense 2
TERM_WORD	décision
mesure	parole
note	connaissance
exemple	engagement
temps	fin
initiative	retraite
part	

Common informant values for each sense

Pr(English | Sense 1)		Pr(English | Sense 2)	
to_take	.433	to_make	.186
to_make	.061	to_speak	.105
to_do	.051	to_rise	.066
to_be	.045	to_take	.066
		to_be	.058
		decision	.036
		to_get	.025
		to_have	.021

Probabilities of English translations

Figure 2: Senses for the French word *prendre*

Figure 2 shows the question that was constructed for the verb *prendre*. The noun to the right yielded the most information, .381 bits, about the English translation of *prendre*. The box in the top of the figure shows the words which most frequently occupy that site, that is, the nouns which appear to the right of *prendre* with a probability greater than one part in fifty. An instance of *prendre* is assigned the first or second sense depending on whether the first noun to the right appears in the left-hand or the right-hand column. So, for ex-

Word: vouloir
Informant: Verb tense
Information: .349 bits

Sense 1	Sense 2
3rd p sing present	1st p sing conditional
1st p sing present	3rd p sing conditional
3rd p plur present	3rd p plur conditional
1st p plur present	3rd p plur subjunctive
2nd p plur present	1st p plur conditional
3rd p sing imperfect	
1st p sing imperfect	
3rd p sing future	

Common informant values for each sense

Pr(English | Sense 1)		Pr(English | Sense 2)	
to_want	.484	to_like	.391
to_mean	.056	to_want	.169
to_be	.056	to_have	.083
to_wish	.033	to_wish	.066
to_refer	.022	me	.029
to_like	.020		

Probabilities of English translations

Figure 3: Senses for the French word *vouloir*

Word: depuis
Informant: Word to the right
Information: .738 bits

Sense 1	Sense 2
longtemps	le
de	la
un	.
quelques	,
deux	l'
1	ce
plus	les
trois	1968

Common informant values for each sense

Pr(English | Sense 1)		Pr(English | Sense 2)	
for	.432	since	.772
last	.123	from	.040
long	.102		
past	.078		
over	.027		
in	.022		
overdue	.021		

Probabilities of English translations

Figure 4: Senses for the French word *depuis*

ample, if the noun to the right of *prendre* is *décision, parole,* or *connaissance,* then *prendre* is assigned the second sense. The box at the bottom of the figure shows the most probable translations of each of the two senses. Notice that the English verb *to_make* is three times as likely when *prendre* has the second sense as when it has the first sense. People *make* decisions, speeches, and acquaintances, they do not *take* them.

Figure 3 shows our results for the verb *vouloir*. Here, the best informant is the tense of *vouloir*. The first sense is three times more likely than the second sense to translate as *to_want*, but twelve times less likely to translate as *to_like*. In polite English, one says *I would like so and so* more commonly than *I would want so and so.*

The question in Figure 4 reduces the entropy of the translation of the French preposition *depuis* by .738 bits. When *depuis* is followed by an article, it translates with probability .772 to *since*, and otherwise only with probability .016.

Finally, consider the English word *cent*. In our text, it is either a denomination of currency, in which case it is usually preceded by a number and translated as *c.*, or it is the second half of *per cent*, in which case it is preceded by *per* and translated along with *per* as *%*. The results in Figure 5 show that the algorithm has discovered this, and in so doing has reduced the entropy of the translation of *cent* by .378 bits.

Word: cent
Informant: Word to the left
Information: .378 bits

Sense 1	Sense 2
per	0
	8
	5
	2
	a
	one
	4
	7

Common informant values for each sense

Pr(French | Sense 1)		Pr(French | Sense 2)	
%	.891	c.	.592
		cent	.239
		sou	.046
		%	.022

Probabilities of French translations

Figure 5: Senses for the English word *cent*

Pleased with these results, we incorporated sense-assignment questions for the 500 most common English words and 200 most common French words into our translation system. This system is an enhanced version of the one described by Brown *et al.* [Brown *et al.*, 1990] in that it uses a trigram language model, and has a French vocabulary of 57,802 words, and an English vocabulary of 40,809 words. We translated 100 randomly selected Hansard sentences each of which is 10 words or less in length. We judged 45 of the resultant translations as acceptable as compared with 37 acceptable translations produced by the same system running without sense-disambiguation questions.

FUTURE WORK

Although our results are promising, this particular method of assigning senses to words is quite limited. It assigns at most two senses to a word, and thus can extract no more than one bit of information about the translation of that word. Since the entropy of the translation of a common word can be as high as five bits, there is reason to hope that using more senses will further improve the performance of our system. Our method asks a single question about a single word of context. We can think of this as the first question in a decision tree which can be extended to additional levels [Lucassen, 1983, Lucassen and Mercer, 1984, Breiman *et al.*, 1984, Bahl *et al.*, 1989]. We are working on these and other improvements and hope to report better results in the future.

REFERENCES

[Bahl *et al.*, 1989] Bahl, L., Brown, P., de Souza, P., and Mercer, R. (1989). A tree-based statistical language model for natural language speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37:1001–1008.

[Breiman *et al.*, 1984] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, California.

[Brown *et al.*, 1990] Brown, P. F., Cocke, J., DellaPietra, S. A., DellaPietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

[Brown *et al.*, 1988] Brown, P. F., Cocke, J., DellaPietra, S. A., DellaPietra, V. J., Jelinek, F., Mercer, R. L., and Roossin, P. S. (1988). A statistical approach to language translation. In *Proceedings of the 12th International Conference on Computational Linguistics*, Budapest, Hungary.

[Brown *et al.*, 1991] Brown, P. F., DellaPietra, S. A., DellaPietra, V. J., and Mercer, R. L. (1991). Parameter estimation for machine translation. In preparation.

[Ide and Véronis, 1990] Ide, N. and Véronis, J. (1990). Mapping dictionaires: A spread-

ing activation approach. In *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, pages 52–64, Waterloo, Canada.

[Lesk, 1986] Lesk, M. E. (1986). Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the SIGDOC Conference.*

[Lucassen, 1983] Lucassen, J. M. (1983). Discovering phonemic baseforms automatically: an information theoretic approach. Technical Report RC 9833, IBM Research Division.

[Lucassen and Mercer, 1984] Lucassen, J. M. and Mercer, R. L. (1984). An information theoretic approach to automatic determination of phonemic baseforms. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 42.5.1–42.5.4, San Diego, California.

[Merialdo, 1990] Merialdo, B. (1990). Tagging text with a probabilistic model. In *Proceedings of the IBM Natural Language ITL*, pages 161–172, Paris, France.

[Nadas *et al.*, 1991] Nadas, A., Nahamoo, D., Picheny, M. A., and Powell, J. (1991). An iterative "flip-flop" approximation of the most informative split in the construction of decision trees. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, Canada.

[White, 1988] White, J. S. (1988). Determination of lexical-semantic relations for multi-lingual terminology structures. In *Relational Models of the Lexicon.* Cambridge University Press, Cambridge, UK.