Using a Target Language Model for Domain Independent Lexical Disambiguation

Jim Cowie, Yevgeny Ludovik, Sergei Nirenburg
Computing Research Laboratory
New Mexico State University
USA
(jcowie, eugene, sergei)@crl.nmsu.edu

Abstract

In this paper we describe a lexical disambiguation algorithm based on a statistical language model we call maximum likelihood disambiguation. The maximum likelihood method depends solely on the target language. The model was trained on a corpus of American English newspaper texts. Its performance was tested using output from a transfer based translation system between Turkish and English. The method is source language independent, and can be used for systems translating from any language into English.

1 Introduction

One of the more persistent problems in machine translation is lexical disambiguation. We propose here a statistics-based solution to the problem in which disambiguation is based entirely on the target language. We created a statistical language model allowing us to compute the probability of any sequence of English words. We assume that any word in a source language may have several different translations depending on its context. Given all the candidate translations of words in the source language, our statistical model finds a word sequence in the target language that is the most probable translation of an input sentence.

In this paper we first outline the problem, describe our approach, and introduce the statistical language model. Then we describe the algorithm for finding the most probable sentence from the set represented by the input ambiguous translation. Finally, we present some experimental results and discuss our future plans.

2 The Problem of Lexical Selection

One source of translation ambiguity is crosslingual polysemy, that is, a situation when a word in the source language has multiple possible translations in the target language. This type of ambiguity is very common, as most source words may be translated into a set of "target language forms" in the target language. This is illustrated by the Spanish-to-English examples in Table 1.

Headword	Translations
presidente	President, prime minister, Presiding judge, presiding magistrate, mayor
cima	top, peak, summit, height

Table 1: Example Translation Equivalents

The transfer component of a typical MT system outputs target language syntactic structures and features plus the translation of open class lexical material. Because of ambiguity (lexical and other), the transfer module overgenerates and outputs many candidate translations of a source sentence. Among these candidates, some may provide a reader with enough information to understand the topic, but generally at best one or two would be considered to be appropriate translations. Current commercial MT systems have difficulty eliminating overgeneration, that is, selecting from a set of possible translations. Phrasal glossaries may help to produce correct translations for collocations, but for individual words the typical method is to select as the translation the most frequent of the target language words in the candidate list. Although this approach is better than a random selection it obviously leads to problems, and results in a poor quality translation.

The following example illustrates the benefits of looking at the co-occurrence properties of words as part of the translation process. The Spanish text is

translated using the first translation found in the lexicon (assumed to be the most frequent translation).

Spanish Source

> En Los Andes peruanos han fallecido por lo menos 6 personas a consecuencia de las fuertes tormentas de nieve que han afectado a la región. Un gran número de autobuses y vehiculos con un total de unas 800 personas permanecen aislados desde el viernes pasado.

Machine Translation into English Using Most Frequent Candidate Method

> In the Andes Peruvian have died at least 6 persons to importance of the strong **storms of ice-cream** that have affected to the region. A great number of buses and vehicles with a total of some 800 persons stay insulated from the last Friday.

Corpus Results

"Nieve" in Latin American Spanish means both *"snow "* and *"ice cream". A* search for combinations of « storm » with « snow » and « ice cream », using a Boolean information retrieval system indexing a large English newspaper corpus (AP, San Jose Mercury, Wall Street Journal, and Financial Times) yields the following results:

> NEWS Concatenated AP, SJM, WSJ, FT
> BRS Search Mode—Enter Query

5_: storm with snow
STORM ..... 7542 docs
SNOW ..... 4803 docs
5_: STORM WITH SNOW.....547 docs

6_: storm with "ice cream"
STORM ..... 7542 docs
"ICECREAM" ..... 5750 docs
6_: "ICE CREAM " WITH STORM .....0 docs

Although both *"snow"* and *"ice cream"* occur in about the same number of documents, the overlap with *"storm"* is zero for the second combination. If the English translation had taken into account the relationship between *"snow"* and *"storm"* in everyday texts, *"storms of ice cream"* would have never occurred. In fact, the translated sentence contains several other less glaring examples of poor lexical selection (to importance − as a consequence, insulated − isolated). Any successful lexical selection method must be able to detect relationships between all the words in a sentence.

3 Improving Lexical Choice

There are several different approaches to improving lexical disambiguation. One can use semantics to detect the concepts associated with the words in the source language and use this to control lexical selection in the target language. This is the approach adopted in the Mikrokosmos project (e.g., Onyshkevych and Nirenburg 1995). The resources required for this approach are very costly to create, which means that current systems relying on this approach are limited to processing texts in a specific domain.

To provide domain independent disambiguation, we decided to experiment with a statistical approach trained on the target language (English). The components of our system, prior to disambiguation, are morphology, lexical lookup, syntactic analysis, syntactic. lexical and feature transfer and English surface form generation. At this point the results consist of a set c sentences reflecting the various ambiguities that have occurred in prior processing—lexical choice, selection of boundaries of syntactic constituents, grammatical features and word and phrase order. For each possible ordering of words and phrases in a target sentence words or phrases with multiple candidate translation are generated as lists with any morphological changes required by the features produced by the previous steps in the system. The statistical model is presented with the most likely set and is required to choose the "best" sentence (best being the most probable sequence of words). We recognize that there are many cases where this method will fail, but we feel that this approach should lead to a significant improvement in lexical selection over the “most frequently occurring” choice.

4 The Target Language Model

4.1 The Problem

The statistical approach suggests that the probability of a word w_n given its left context is defined by the conditional probability distribution $p(w_n | w_1 w_2 ... w_{n-1})$ The probability of any sequence of words $\{w_1, w_2, \ldots w_N\}$ can be computed as a product of conditional probabilities of every word given its left context:

$$p(w_1, w_2, \ldots, w_N) = \prod_{n=1}^{N} p(w_n \mid w_1, \ldots, w_{n-1}).$$

If the input data is represented by a sequence of sets of possible translations $\{W_1\ W_2.\ \ldots,\ W_N\}$, then the disambiguation task is reduced to finding the most probable sequence of words $w_n^* \in W_n$:

$$\{w_1^*,\ldots,w_N^*\} = \arg\max_{(w_1 \in W_1,\ldots,w_N \in W_N)} p(w_1,\ldots,w_N).$$

4.2 The Statistical Language Model

It is obvious that the longer a context, the better the model. However even tri-gram models contain too many parameters, and many tri-grams will not be attested in a corpus, so that it will be impossible to estimate their probabilities. The case for four-grams will be even more problematic. Our language model tries to take into account contexts of 5 words, but all dependencies are approximated using the concept of a distant bi-gram (Huang et al. 1992).

An i-distant bi-gram (w0, wl) is a pair of words that occurred in a sentence so that the position of the word w0 was the position of the word wl minus i. Thus 1-distant bi-gram is a traditional bi-gram consisting of two consecutive words. Our statistical model makes use of 5 i-distant bi-grams, 1 <= i <= 5, and can be represented by the following equation:

$$p(w_n \mid w_1,\ldots,w_{n-1}) = p(w_n \mid w_{n-4},\ldots,w_{n-1}) = \sum_{i=1}^{5} \lambda_i * p(w_n \mid w_{n-i}).$$

This approach allows us to use longer-distance dependencies than a tri-gram model. However, as can be expected, there is a price to pay: tri-gram models are much more efficient in representing dependencies between three consecutive words. In the equation above, the parameters $\{\lambda_i\}$ must be set to sum up to 1.

To train our model we used the Wall Street Journal for 1987 and 1988, and the San Jose Mercury for 1995, as provided for the TREC information retrieval evaluations (Harman, 1996). These contain around 73 million words. First we processed the corpus data deleting SGML-tags and punctuation marks. We kept all sentences separate and represented all consecutive digits in the text as a single digit, always the same one. Then we computed the frequencies of all observed i-distant bi-grams, 1<= i <=5, and frequencies of all words, and finally all conditional probabilities dividing the frequency of an i-distant bigram by the frequency of a condition word.

4.3 The Most Probable Word Sequence Search Algorithm

The algorithm implements the k-best search, and can be described recursively. Suppose we have *k* best word sequences of length n: $w(j)=\{w_1(j).....w_n(j)\}$, $1 \le j \le k$, together with probabilities $p_n(j)=p(w_1(j).....w_n(j))$. If *C(j)* is a set of words, each of which can immediately follow The word sequence *w(j)*, we create all sequences *w(j)xw*, $w \in C(j)$ for all *j*, $1 \le j \le k$ and compute their probabilities according to the following formula:

$$p(w_1(j), .. w_n(j),w) = p_n(j) * p(w \mid w_{n-5}(j),..w_{n-i}(j))$$

$w \in C(j)$

Then we pick up *k* best, that is, most probable of them.

Although the discussion has been phrased in terms of word alternative disambiguation, those alternatives can be phrases or empty words, so the sentences that are ranked can have different length.

5 Evaluation of the Approach

We evaluated the approach on the results of a Turkish-English MT. The input data was created by the system described in Beale (1995).

Here is one of the sentences used for evaluation:

```
("Privatization" "special") of public-
sector shares that was in transfer in
("cellular" "pocket") telephones of
*gsm* to ("private" "special") sector
and Turkey ("Business" "Job") Bank is
one important step that was ("taken"
"thrown") in ("privatization" "spe-
cial") area in Turkey
```

Words in parentheses are alternative translations.

The evaluation results based on six sentences obtained through the Turkish-English MT system are represented in the table below. We compared the sentences produced by our program with versions selected by a native speaker of English. For the above ambiguous translation, the manually produced version was:

```
("Privatization") of public-sector
shares that was in transfer in ("cel-
lular") telephones of *gsm* to ("pri-
vate") sector and Turkey ("Business")
Bank is one important  step that was
("taken") in ("privatization") area in
Turkey
```

The output of the 5 i-distant bi-grams version of our program was:

```
("Privatization") of public-sector
shares that was in transfer in ("cel-
lular") telephones of *gsm* to ("pri-
vate") sector and Turkey ("Business")
```

```
Bank is one important step that was
("taken") in ("special") area in Tur-
key
```

If an alternative word translation chosen by the program was different from the one selected by the human expert, the error rate was incremented by 1. The percent error rate was computed by dividing the error rate by the total number of choice points in all sentences (22). Out of the 22 ambiguous positions, 19 were selections from among two alternative translations, the remainder were selections from among three alternatives.

The evaluation results for the six sentences are represented in the Table 2. The results produced by the Unigram method are equivalent to choosing the most frequently occurring use in the corpus, and so performance here is what might be expected of the current MT systems. Only a small improvement is found for bi-grams implying that collocation is probably less powerful as a disambiguation technique than one might expect.

Table 2: Evaluation Results

	Random choice	*Uni-grams*	*Bi-grams*	*5 bi-grams*
% error rate	53.0	27.3	22.7	13.6

6. Future Plans

The system has also been used on outputs from fast ramp-up systems of Japanese and Korean to English MT. In these cases the quality of the lexicon and morphology was not sufficiently high to produce good throughput. We are working on tuning the lexicon to improve the word choice and relaxing the results of morphology to allow more choices to be processed by later stages of the system. Our results at present are based on a very small sample, and we intend to extend this to test corpora of at least 100 sentences per language as soon as outputs from the corresponding MT systems will become available. We are also planning to develop an i-trigram model, which will operate similarly to the i-bigram model described here. An enhanced version of the i-bigram approach that will allow us to have an option to add word alternatives missing in the original input data is also being investigated.

Acknowledgements

The work described here has depends on the output of many components developed by our colleagues at the Computing Research Laboratory. Thanks, in particular, are due to Dr. Stephen Beale for producing the ambiguous throughput used in our experiments. The research was partially funded by DoD contract MDA904-97-C-3076.

References

Beale, S., S. Nirenburg, J. Cowie and K. Oflazer. 1999. Quick Ramp-Up MT: The Pin-the-tail-on-the-Donkey Approach. Internal Memo. NMSU CRL.

BRS/Search Users Guide (1995) Dataware Technologies.

Cowie, J., Guthrie, J., and Guthrie, L. (1992) “Lexical Disambiguation using Simulated Annealing”, Computing Research Laboratory, Las Cruces NM, Fifth DARPA Workshop on Speech & Natural Language. 1992.

Gale, W.A., Church, K.W., Yarowsky, D. (1992) “One Sense Per Discourse”, Fifth DARPA Workshop on Speech & Natural Language.

Harman, D. ed. (1996) The Fifth Text Retrieval Conference (TREC-5), Gaithersburg, MD: Computer Systems Laboratory, NIST.

Guthrie, L., Guthrie, J., and Cowie, J. (1994) “Resolving Lexical Ambiguity”, in Corpus Based Research into Language, Rodopi, Amsterdam-Atlanta GA, 1994

Huang, X. D., Alleva, F., Hon, H. W., Hwang, M. Y., Lee, K.F. and Rosenfeld, R. (1992) “The SPHINX-II Speech Recognition System: An Overview”, in Computer Speech and Language

Onyskevych, B. and Nirenburg, S. (1995) “A Lexicon for Knowledge-Based MT”, Machine Translation, 10:1-2

Vanni, M. and Zajac, R. (1996) “The Temple Translator's Workstation Project” in Proceedings of the Tipster-II 24-month Workshop. Tysons Corner, VA: DARPA