Cognates Can Improve Statistical Translation Models

Grzegorz Kondrak
Department of Computing Science
University of Alberta
221 Athabasca Hall
Edmonton, AB, Canada T6G 2E8
kondrak@cs.ualberta.edu

Daniel Marcu and **Kevin Knight**
Information Sciences Institute
University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA, 90292
marcu,knight@isi.edu

Abstract

We report results of experiments aimed at improving the translation quality by incorporating the cognate information into translation models. The results confirm that the cognate identification approach can improve the quality of word alignment in bitexts without the need for extra resources.

1 Introduction

In the context of machine translation, the term *cognates* denotes words in different languages that are similar in their orthographic or phonetic form and are possible translations of each other. The similarity is usually due either to a genetic relationship (e.g. English *night* and German *nacht*) or borrowing from one language to another (e.g. English *sprint* and Japanese *supurinto*). In a broad sense, cognates include not only genetically related words and borrowings but also names, numbers, and punctuation. Practically all bitexts (bilingual parallel corpora) contain some kind of cognates. If the languages are represented in different scripts, a phonetic transcription or transliteration of one or both parts of the bitext is a pre-requisite for identifying cognates.

Cognates have been employed for a number of bitext-related tasks, including sentence alignment (Simard et al., 1992), inducing translation lexicons (Mann and Yarowsky, 2001), and improving statistical machine translation models (Al-Onaizan et al., 1999). Cognates are particularly useful when machine-readable bilingual dictionaries are not available. Al-Onaizan et al. (1999) experimented with using bilingual dictionaries and cognates in the training of Czech–English translation models. They found that appending probable cognates to the training bitext significantly lowered the perplexity score on the test bitext (in some cases more than when using a bilingual dictionary), and observed improvement in word alignments of test sentences.

In this paper, we investigate the problem of incorporating the potentially valuable cognate information into the translation models of Brown et al. (1990), which, in their original formulation, consider lexical items in abstraction of their form. For training of the models, we use the GIZA program (Al-Onaizan et al., 1999). A list of likely cognate pairs is extracted from the training corpus on the basis of orthographic similarity, and appended to the corpus itself. The objective is to reinforce the co-ocurrence count between cognates in addition to already existing co-ocurrences. The results of experiments conducted on a variety of bitexts show that cognate identification can improve word alignments, which leads to better translation models, and, consequently, translations of higher quality. The improvement is achieved without modifying the statistical training algorithm.

2 The method

We experimented with three word similarity measures: Simard's condition, Dice's coefficient, and LCSR. Simard et al. (1992) proposed a simple condition for detecting probable cognates in French–English bitexts: two words are considered cognates if they are at least four characters long and their first four characters are identical. Dice's coefficient is defined as the ratio of the number of shared character bigrams to the total number of bigrams in both words. For example, *colour* and *couleur* share three bigrams (*co*, *ou*, and *ur*), so their Dice's coefficient is $\frac{6}{11} \simeq 0.55$. The Longest Common Subsequence Ratio (LCSR) of two words is computed by dividing the length of their longest common subsequence by the length of the longer word. For example, LCSR(*colour*,*couleur*) = $\frac{5}{7} \simeq 0.71$, as their longest common subsequence is "c-o-l-u-r".

In order to identify a set of likely cognates in a tokenized and sentence-aligned bitext, each aligned segment is split into words, and all possible word pairings are stored in a file. Numbers and punctuation are not considered, since we feel that they warrant a more specific approach. After sorting and removing duplicates, the file represents all possible one-to-one word alignments of the bitext. Also removed are the pairs that include English function words, and words shorter than the minimum length (usually set at four characters). For each word pair, a similarity measure is computed, and the file is again sorted, this time by the computed similarity value. If the measure returns a non-binary similarity value, true cognates are very frequent near the top of the list, and become less frequent towards the bottom. The set of likely cognates is obtained by selecting all pairs with similarity above a certain threshold. Typically, lowering the threshold increases recall while decreasing precision of the set. Finally, one or more copies of the resulting set of likely cognates are concatenated with the training set.

3 Experiments

We induced translation models using IBM Model 4 (Brown et al., 1990) with the GIZA toolkit (Al-Onaizan et al., 1999). The maximum sentence length in the training data was set at 30 words. The actual translations were produced with a greedy decoder (Germann et al., 2001). For the evaluation of translation quality, we used the BLEU metric (Papineni et al., 2002), which measures the n-gram overlap between the translated output and one or more reference translations. In our experiments, we used only one reference translation.

3.1 Word alignment quality

In order to directly measure the influence of the added cognate information on the word alignment quality, we performed a single experiment using a set of 500 manually aligned sentences from Hansards (Och and Ney, 2000). Giza was first trained on 50,000 sentences from Hansards, and then on the same training set augmented with a set of cognates. The set consisted of two copies of a list produced by applying the threshold of 0.58 to LCSR list. The duplication factor was arbitrarily selected on the basis of earlier experiments with a different training and test set taken from Hansards.

The incorporation of the cognate information resulted in a 10% reduction of the word alignment error rate, from 17.6% to 15.8%, and a corresponding improvement in both precision and recall. An examination of randomly selected alignments confirms the observation of Al-Onaizan et al. (1999) that the use of cognate information reduces the tendency of rare words to align to many co-occurring words.

In another experiment, we concentrated on co-occurring identical words, which are extremely likely to represent mutual translations. In the baseline model, links were induced between 93.6% of identical words. In the cognate-augmented model, the ratio rose to 97.2%.

3.2 Europarl

Europarl is a tokenized and sentence-aligned multilingual corpus extracted from the Proceedings of the European Parliament (Koehn, 2002). The eleven official European Union languages are represented in the corpus. We consider the variety of languages as important for a validation of the cognate-based approach as general, rather than language-specific.

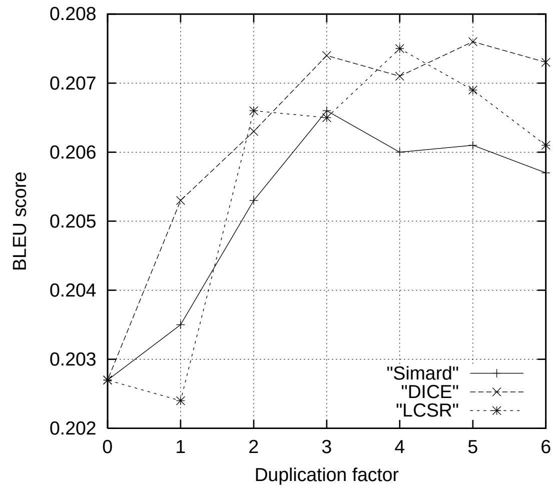

Figure 1: BLEU scores as a function of the duplication factor for five methods of cognates identification averaged over nine language pairs.

As the training data, we arbitrarily selected a subset of the corpus that consisted the proceedings from October 1998. By pairing English with the remaining languages, we obtained nine bitexts[1], each comprising about 20,000 aligned sentences (500,000 words). The test data consisted of 1755 unseen sentences varying in length from 5 to 15 words from the 2000 proceedings (Koehn, 2002). The English language model was trained separately on a larger set of 700,000 sentences from the 1996 proceedings.

Figure 1 shows the BLEU scores as a function of the duplication factor for three methods of cognates identification averaged over nine language pairs. The results averaged over a number of language pairs are more informative than results obtained on a single language pair, especially since the BLEU metric is only a rough approximation of the translation quality, and exhibits considerable variance. Three different similarity measures were compared: Simard, DICE with a threshold of 0.39, and LCSR with a threshold of 0.58. In addition, we experimented with two different methods of extending the training set with with a list of cognates: one pair as one sentence (Simard), and thirty pairs as one sentence (DICE and LCSR).[2]

[1] Greek was excluded because its non-Latin script requires a different type of approach to cognate identification.

[2] In the vast majority of the sentences, the alignment links are correctly induced between the respective cognates when multi-

Threshold	Pairs	Score
Baseline	0	0.2027
0.99	863	0.2016
0.71	2835	0.2030
0.58	5339	0.2058
0.51	7343	0.2073
0.49	14115	0.2059

Table 1: The number of extracted word pairs as a function of the LCSR threshold, and the corresponding BLEU scores, averaged over nine Europarl bitexts.

The results show a statistically significant improvement[3] in the average BLEU score when the duplication factor is greater than 1, but no clear trend can be discerned for larger factors. There does not seem to be much difference between various methods of cognate identification.

Table 1 shows results of augmenting the training set with different sets of cognates determined using LCSR. A threshold of 0.99 implies that only identical word pairs are admitted as cognates. The words pairs with LCSR around 0.5 are more likely than not to be unrelated. In each case two copies of the cognate list were used. The somewhat surprising result was that adding only "high confidence" cognates is less effective than adding lots of dubious cognates. In that particular set of tests, adding only identical word pairs, which almost always are mutual translations, actually decreased the BLEU score. Our results are consistent with the results of Al-Onaizan et al. (1999), who observed perplexity improvement even when "extremely low" thresholds were used. It seems that the robust statistical training algorithm has the ability of ignoring the unrelated word pairs, while at the same time utilizing the information provided by the true cognates.

3.3 A manual evaluation

In order to confirm that the higher BLEU scores reflect higher translation quality, we performed a manual evaluation of a set of a hundred six-token sentences. The models were induced on a 25,000 sentences portion of Hansards. The training set was augmented with two copies of a cognate list obtained by thresholding LCSR at 0.56. Results of a manual evaluation of the entire set of 100 sentences are shown in Table 2. Although the overall translation quality is low due to the small size of the training corpus and the lack of parameter tuning, the number of completely acceptable translations is higher when cognates are added.

Evaluation	Baseline	Cognates
Completely correct	16	21
Syntactically correct	8	7
Semantically correct	14	12
Wrong	62	60
Total	100	100

Table 2: A manual evaluation of the translations generated by the baseline and the cognate-augmented models.

4 Conclusion

Our experimental results show that the incorporation of cognate information can improve the quality of word alignments, which in turn result in better translations, In our experiments, the improvement, although statistically significant, is relatively small, which can be attributed to the relative crudeness of the approach based on appending the cognate pairs directly to the training data. In the future, we plan to develop a method of incorporating the cognate information directly into the training algorithm. We foresee that the performance of such a method will also depend on using more sophisticated word similarity measures.

ple pairs per sentence are added.

[3]Statistical significance was estimated in the following way. The variance of the BLEU score was approximated by randomly picking a sample of translated sentences from the test set. The size of the test sample was equal to the size of the test set (1755 sentences). The score was computed in this way 200 times for each language. The mean and the variance of the nine-language average was computed by randomly picking one of the 200 scores for each language and computing the average. The mean result produced was 0.2025, which is very close to the baseline average score of 0.2027. The standard deviation of the average was estimated to be 0.0018, which implies that averages above 0.2054 are statistically significant at the 0.95 level.

References

Y. Al-Onaizan, J. Curin, M. Jahr, K. Knight, J. Lafferty, D. Melamed, F. Och, D. Purdy, N. Smith, and D. Yarowsky. 1999. Statistical machine translation. Technical report, Johns Hopkins University.

P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. 1990. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

U. Germann, M. Jahr, K. Knight, D. Marcu, and K. Yamada. 2001. Fast decoding and optimal decoding for machine translation. In *Proceedings of ACL-01*.

P. Koehn. 2002. Europarl: A multilingual corpus for evaluation of machine translation. In preparation.

G. Mann and D. Yarowsky. Multipath translation lexicon induction via bridge languages. In *Proceedings of NAACL 2001*.

F. J. Och and H. Ney. 2000. Improved statistical alignment models. In *Proceedings of ACL-00*.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL-02*.

M. Simard, G. F. Foster, and P. Isabelle. 1992. Using cognates to align sentences in bilingual corpora. In *Proceedings of TMI-92*.