Learning Translations of Named-Entity Phrases from Parallel Corpora*

Robert C. Moore
Microsoft Research
Redmond, WA 98052, USA
bobmoore@microsoft.com

Abstract

We develop a new approach to learning phrase translations from parallel corpora, and show that it performs with very high coverage and accuracy in choosing French translations of English named-entity phrases in a test corpus of software manuals. Analysis of a subset of our results suggests that the method should also perform well on more general phrase translation tasks.

1 Introduction

Machine translation can benefit greatly from augmenting knowledge of word translations with knowledge of phrase translations. Multiword phrases may have nonliteral translations, or one of several equally valid literal translations may be strongly preferred in practice. Automatically learning translations of single words from parallel corpora has been much studied over the past ten years or so (Melamed, 2000, and references), but learning translations of multiword phrases has received less attention. (See Section 5 for a review of prior work in this area.) In this paper, we develop a new approach to learning phrase translations from parallel corpora, and show that it performs with very high coverage and accuracy on a named-entity phrase translation task. Moreover, analysis of a subset of our evaluation results suggests that the method should also perform well on more general phrase translation tasks.

In our approach, we are given a sentence-aligned parallel corpus annotated with a set of phrases in one of the two languages (the *source language*), and our goal is identify the corresponding phrases in the corpus in the other language (the *target language*), ranking the translation pairs in order of confidence. Certain segments of the target language corpus may be annotated as constituting lexical compounds, which may or may not include the translations of the source language phrases of interest. Otherwise there is no annotation of the target language text, except for its being divided into words and sentences.

Below we describe the issues in named-entity phrase translation motivating this research, we explain our algorithm, and we present the results of our evaluation on a named-entity phrase translation task. We pay particular attention to the subset of the data that lacks the special characteristics of the named-entity task that we take advantage of to optimize our performance, to suggest how the algorithm might perform on more general tasks. Finally we compare our approach and results to previous work on learning phrase translations.

2 The Named-Entity Phrase Translation Task

Named-entity expressions (Chinchor and Marsh, 1997) are any words or phrases that name a specific entity. While often thought of in terms of categories such as persons, organizations, or locations, in technical text a much wider range of

*Revised version of paper appearing in *Proceedings of EACL 2003*, Budapest, Hungary. Revised 1 May 2003.

types of entities are often named. In software manuals, for example, named-entity expressions include names of menu items, dialogue boxes, software systems, etc. While named-entity expressions are typically used as proper nouns, those encountered in technical text often do not have the syntactic form of nouns or noun phrases. Consider, *Click the View Source Tables button*. In this sentence, *View Source Tables* has the syntactic form of a nonfinite verb phrase, but it is used like a proper noun. It would be difficult to recognize as a named-entity expression, except for the fact that in English, all or most of the words in named-entity expressions are typically capitalized.

Capitalization conventions of French and Spanish, however, make it harder to recognize named-entity phrases, because often only the first word of the phrase is capitalized. For example, in our data, the French translation of *View Source Tables* is *Afficher les tables source*. Embedded in a sentence, it is difficult to determine the extent of such a named-entity expression using only monolingual lexical information. If we could fully parse the sentence, we might be able to recognize *Afficher les tables source* as a named-entity expression; but it is very difficult to parse a sentence where something that looks like a nonfinite verb phrase is used like a proper noun, unless the parser already knows that there is something special about that phrase. Our problem, therefore, is to find the phrases that are translations of the English expressions, without necessarily having previously recognized that they are in fact complete phrases.

Our approach addresses the identification and translation problems simultaneously. Taking English as our source language, we use capitalization clues to identify named-entity phrases in the English portion of a sentence-aligned parallel corpus, and then we apply statistical techniques to decide which contiguous sequences of words in the target language portion of the corpus are most likely to correspond to the English phrases. We can then add the learned named-entity phrases to a phrasal lexicon that can be used to better parse target language sentences, as well as adding the translation pairs to a bilingual translation dictionary.

3 The Algorithm

Our algorithm begins by computing a fairly simple bilingual-word-association metric on the parallel corpus, which is then used to construct three progressively more refined phrase translation models. The first model is used only to initialize the second, which in turn is used only to initialize the third, which is the model actually used. Although the algorithm is designed to take advantage of some special properities of named-entity phrase translation, it is in no way limited to this task, and can be applied to any phrase translation task in which a set of fixed phrases can be indentified on one side of a bilingual parallel corpus, whose translations on the on the other side are desired. A random sample of the output of our phrase translation learner is shown in Table 1.[1] All these examples, except for the last, were judged to be correct in context in our evaluation.

3.1 Model 1

In addition to statistics derived from the corpus, the first model embodies two nonstatistical heuristics. The first is simply that we do not hypothesize translations of source language phrases that would require splitting predetermined lexical compounds, if any, in the target language. The second heuristic is that if the phrase whose translation is sought occurs in exactly the same form in the target language sentence as in the source language sentence, we assume that it is the corresponding phrase in that sentence with probability 1.0. This is a very important heuristic in our test corpus, because almost 17% of the source language test phrases are names or technical terms that occur untranslated in the target language text.

We start by measuring the degree of association between a source language word s and a target language word t (ignoring upper/lower case distinctions) in terms of the frequencies with which s occurs in sentences of the source language part of the corpus and t occurs in sentences of the target language part of the corpus, compared to the frequency with which s and t co-occur in aligned sentences of the corpus. The particular measure

[1] Words joined by "_" were indentified as compounds by the monolingual tokenizers prior to applying our algorithm.

MSMQ_Explorer	explorateur MSMQ
Highlighted_Edges	Contours en surbrillance
ADD_FILEGROUP	ADD FILEGROUP
Custom_Preview_Area	Aperçu personnalisé
Web_Proxy_Server	serveur proxy Web
Windows_NT_3.5	Windows NT version 3.5
All_Unassigned	Tous non assignés
Build_Query	Générer la requête
Product_Support_Services_Web	Web_des_Services_de_Support_Technique
Microphone_Settings_Wizard	Assistant_Paramètres de le microphone
Process_Accounting	comptabilisation de les processus
SQL-DMO_Examples	Exemples SQL_-_DMO
Flexible_Data_Model	Modèle de données flexible
SQL_Server_Log_Reader_Agent	agent de lecture de le journal SQL_Server
NT_LM_Security_Support_Provider	Fournisseur de le service de sécurité NT_LM
Flip_On_Short_Edge	Retourner sur les bords courts
Transact_SQL	Transact_SQL
Microsoft_Repository	Registre de stockage de Microsoft
Sort_Orders	ordres de tri
Microsoft_Distributed_Transaction_Coordinator	transaction distribuée

Table 1: Random sample of translations produced.

we use is the log-likelihood-ratio statistic recommended by Dunning (1993).

In the past we have found that this word-association metric provides an excellent basis for learning single-word translation relationships, and the higher the score, the more likely the association is to be a true translation relation. However, with this particular metric there is no obvious way to combine the scores for indvidual word pairs into a composite score for phrase translation candidates; so we use the scores indirectly to estimate some relevant probabilities, which can then be combined to yield a composite score. To do this, we make another pass through the parallel corpus, and for each word s in the source language sentence of an aligned sentence pair, we note which word t in the target language sentence of the pair has the strongest association with s. If there is no word having a positive association with s above a certain cut-off, we take the empty word ϵ to have the highest association with s in the given sentence pair. We do this in both directions, since even if the word most strongly associated with s is t, the word most strongly associated with t might be some other word s'. For each pair of words s and t, we keep a count of how many times t occurs as the word most strongly associated with s, and vice versa. From these counts, we estimate (using a modified form of Good-Turing smoothing) the probability $P_1(t|s)$ that an occurrence of a source language word s will have a word t as its most strongly associated word in the corresponding aligned target language sentence, as well as the probability $P_1'(s|t)$ that an occurrence of a target language word t will have a word s as its most strongly associated word in the corresponding aligned source language sentence.

The key idea of our first model is that if a candidate substring of a target language sentence corresponds to a selected source language phrase, then the words in the candidate target language substring should associate most strongly with words of the selected target language phrase, and the words of the target language sentence outside the candidate substring should associate most strongly with words of the source language sentence outside the selected phrase. We compute a composite score for a particular partitioning of the target

language sentence by summing the logarithms of the association probabilities for the strongest associations we can find of words in the selected source language phrase to words in the candidate target language substring (and vice versa), which we call the *inside score*, added to the sum of the logarithms of the association probabilities for the strongest associations we can find for the words of the source language sentence outside the selected phrase to the words of the target language sentence outside the candidate substring (and vice versa), which we call the *outside score*.

Symbolically, let s, s' be words in the source language sentence S; let t, t' be words in the target language sentence T; let S' be a substring of S; let T' be a substring of T conjectured to be the translation of S'. Then,

$$\begin{aligned}&\mathrm{inside}(S', T') = \\ &\quad \sum_{s \in S'} \max_{t' \in T' \cup \{\epsilon\}} \log(P_1(t'|s)) \; + \\ &\quad \sum_{t \in T'} \max_{s' \in S' \cup \{\epsilon\}} \log(P_1'(s'|t))\end{aligned}$$

$$\begin{aligned}&\mathrm{outside}(S', T') = \\ &\quad \sum_{s \in S - S'} \max_{t' \in (T - T') \cup \{\epsilon\}} \log(P_1(t'|s)) \; + \\ &\quad \sum_{t \in T - T'} \max_{s' \in (S - S') \cup \{\epsilon\}} \log(P_1'(s'|t))\end{aligned}$$

Thus if a target language word outside the candidate translation has a high probability of associating with a source language word in the selected phrase, that candidate translation is likely to get a lower composite score than another candidate translation that does include that particular target language word. While this is not actually a generative model, the probabilities being combined are comparable, and it seems to work well in practice.

Since in named-entity translation from English to Spanish or French, capitalization is relevant in determining the phrase translation (and since the word-association statistic ignores capitalization), we add to the composite score a log probability estimated for three capitalization possibilities: the target language phrase begins with a capitalized word, the target language phrase has no capitalized words, or the target language phrase contains capitalized words, but does not begin with one. Let $P_{\mathrm{capt}}(T')$ represent the probability that a target language translation of a source language named-entity expression falls into the capitalization class of T'. The final expression for the Model 1 score of a source language phrase S' and a hypothesized target language translation T' is, then,

$$\mathrm{outside}(S', T') + \mathrm{inside}(S', T') + \log(P_{\mathrm{capt}}(T'))$$

The capitalization class probabilities are initially taken to be uniform and are iteratively recomputed by Viterbi re-estimation. In this way, we are able to learn that an English named-entity phrase is likely to correspond to a Spanish or French phrase in which the first word is capitalized. This is only a strong probability and not a certainty, however. In the random sample of the output of our system that we selected for evaluation, we found that 20% of the source language phrases had hypothesized target language translations in which the first word is not capitalized.

3.2 Model 2

Model 2 replaces the inside score and capitalization log probability of the Model 1 by a new inside score computed as the logarithm of a holistic estimate of the conditional probability of the target language candidate occurring as the translation of the source language phrase, $P_2(T'|S')$, times the conditional probability of the source language phrase occuring as the translation of the target language candidate, $P_2'(S'|T')$. This unusual statistic was chosen to mirror as closely as possible the structure of the first model; we are simply replacing approximations of these probabilities estimated from sets of single-word associations with estimates based on occurrences of the complete phrases.

This whole-phrase-based inside score is combined with the original word-association-based outside score, using a scale factor α to account for the fact that the new version of the inside score can be expected to have a different degree of variability from the one it is replacing. If we did not do this, the exaggerated variance due to false independence assumptions in the individual probabilities combined in the computation of the outside score would overwhelm the reduced variance of the inside score. The scale factor α is simply the

ratio of the standard deviation of the inside scores as estimated in the first model and the standard deviation of the initial estimates of the inside scores for the second model. The Model 2 scores, then, are of the form

$$\text{outside}(S', T') + \alpha \log(P_2(T'|S') \cdot P_2'(S'|T'))$$

The initial values for the phrase translation probabilities are estimated according to the first model, and iteratively re-estimated using EM, by treating the Model 2 scores as log probabilities and normalizing them across the candidate translations in each sentence pair for each source language phrase.

The effect of moving from Model 1 to Model 2 is to let tendencies in the translation of particular phrases across sentences influence the choice of a translation in a particular sentence. If a given phrase has a clearly preferred translation in several sentences, that can be taken into account in choosing a translation for the phrase in a sentence where the individual word association probabilities leave the translation of the phrase unclear.

3.3 Model 3

Model 3 consists of computing the log-likelihood-ratio metric for all the selected phrases and candidate translations, based on the whole phrases rather than the individual words composing them, but counting as co-occurrences only pairs consisting of a selected phrase and its highest scoring candidate translation in a particular aligned sentence pair. We initialize this model by finding the highest scoring translation of each occurrence of each selected source language phrase according to Model 2, and we iteratively recompute the parameters using Viterbi re-estimation. When this re-estimation converges, we have our final set of phrase translation scores, in terms of the log-likelihood-ratio metric for whole phrases.

The main point of Model 3 is to obtain a consistent set of log-likelihood-ratio scores to use as a confidence measure for the phrase translation pairs. This could be computed just in a single pass, but the Viterbi re-estimation ensures that the data we are computing the log-likelihood-ratio scores from is consistent with the resulting scores. That is, it ensures that we do not count an instance in the data of a particular translation pair, when there is a higher scoring possibility according to the confidence measure we are computing.

4 Evaluation Results

The algorithm was developed using English-Spanish parallel data, and independently tested on 192,711 English-French parallel sentence pairs consisting mainly of computer software manuals. 73,108 occurrences of 12,301 unique multi-word named-entity phrases were hypothesized in the English data by a hand-built rule-based tagger, mainly using capitalization clues.

We evaluated the performance of our algorithm in finding translations for the hypothesized named-entity phrases using a random sample of 1195 of the proposed translations. The correctness of the correspondence between the English phrases and their hypothesized translations was judged by a fluent French-English bilingual, with the aid of the sentence pair for which each hypothesized translation received the highest score, according to Model 1. (In preliminary work, we found that it was very difficult to judge correctness without seeing relevant examples from the data.) In some cases, the existence of words in the French not corresponding to anything in the English led to multiple equally valid phrase correspondences, any of which was judged correct. Clear cases of partial matches, however, were always counted as incorrect.

The results of the evaluation are shown in Table 2. "Cumulative Coverage" means the proportion of the unique phrases for which at least one translation is proposed, proceeding in order of strength of association from highest to lowest. "Cumulative Accuracy" is the estimated accuracy of the translations proposed for the top scoring fraction of translations corresponding to "Cumulative Coverage".[2] "'Good Input' Cumulative Accuracy" is the same as "Cumulative Accuracy", but removing 157 cases (13% of the test data) where

[2]These are essentially the same measures used by Melamed (2000) in his work on learning single-word translations from parallel corpora. We use the coverage metric rather than recall, because in this data, phrases often have more than one translation, and we have no practical way of knowing what proportion of these translations we find. Accuracy is the same as precision.

Cumulative Coverage	All Data			"Hard" Data
	Cumulative Accuracy	"Good Input" Cumulative Accuracy	Singleton Proportion	"Good Input" Cumulative Accuracy
0.100	0.914	0.980	0.000	0.96
0.200	0.906	0.979	0.000	0.87
0.300	0.896	0.975	0.000	0.92
0.400	0.873	0.965	0.087	0.89
0.500	0.879	0.963	0.243	0.90
0.600	0.875	0.961	0.354	0.88
0.700	0.880	0.961	0.436	0.88
0.800	0.870	0.955	0.498	0.86
0.900	0.856	0.941	0.565	0.86
0.950	0.843	0.938	0.595	0.85
0.990	0.808	0.916	0.619	0.84

Table 2: Performance of phrase translation learning algorithm.

it was impossible choose a correct French translation for the English phrase, within the assumptions of the task.[3] "Singleton Proportion" records the proportion of the English test phrases that had only a single occurrence in the data.

These results show accuracy over 80% up to 99% coverage, with accuracy over 91% at 99% coverage when only clean input data is considered. Moreover, at this level 62% of the English phrases had only a single occurrence in the data. Accuracy is very high compared to previous work on phrase translation, but this task does have several properties that probably make it easier than a more general phrase translation task would be. First, 17% of the English phrases were repeated exactly in the French corpus. Second, 14% of the non-identical hypothesized French translations were already identified as complete lexical compounds. Finally, 74% of the hypothesized French translations began with a capital letter.

To test the robustness of our technique to phrase translation learning tasks where these advantages are lacking, we analyzed our evaluation data to find all cases where the tokenizations were correct, but the correct translation of the English phrase began with a lower case letter, and the translation itself was not identified as a lexical compound in preprocessing. (This also guaranteed that none of the translations was identical to the English phrase, since all the English test phrases began with a capital letter.) There were 240 such cases out of our sample of 1195 hypothesized translation pairs. The performance of the algorithm on this "hard" subset of the data is shown in the last column of Figure 2. Compared with the results in the third column on all the "good input" data, the error rates go up by a factor of 2–3, but accuracy is still a quite respectable 84% at 99% coverage.[4]

5 Comparison with Previous Work

Our work on learning phrase translations can be classified along at least two dimensions. First, our approach is asymmetrical in that it assumes that

[3] 85% of these cases were errors (or at least inconsistencies) in identification of lexical compounds in English and/or French that made it impossible to correctly identify the correct French translation of an English phrase. (Smadja et al. [1996] similarly report the performance of their collocation-translation learner, removing errors due to mistakes in identifying the source language collocations.) These included cases where English words were incorrectly included in or omitted from the phrase so that there was no single corresponding French phrase, or where an incorrect identification of a French lexical compound connected words in the translation of the English phrase with words not in the translation. The remaining 15% of the cases excluded from "good input" were cases where the French sentence simply did not contain any phrase corresponding to the English phrase, either because of free translation or because of errors in sentence alignment.

[4] "Cummulative coverage" in this case means coverage of the 235 English phrases that were determined to have at least one lowercase translation.

a set of phrases in the source language is given, and the task is to find their translations in the target language, for which only minimal monolingual knowledge may be available. In symmetrical approaches, the problem is generally viewed as discovering phrases in both languages that are mutual translations, for which equally rich (or equally poor) analysis tools are available. Second, our approach applies only to fixed phrases, since it assumes that the translation of a source language phrase is a contiguous sequence of words in the target language. At least one other reported approach applies to more flexible collocations.

Al-Onaizan and Knight's (2002) work is both asymmetrical and targeted at fixed phrases, as well as being perhaps the only other work aimed specifically at named-entity phrase translation (for Arabic to English). Lacking a parallel bilingual corpus, their methods are completly different from ours, and their reported accuracy is only 65–73%.

Dagan and Church's (1997) *Termight* is also asymmetrical and targeted at fixed phrases. It is conceived of as an automated assistant for a lexicographer that proposes technical terms extracted from a corpus using monolingual methods, and for those approved by the user, proposes possible translations from a parallel corpus. While apparently never intended for use as a fully automatic translation finder, its accuracy if used as such was reported by Dagan and Church to be 40% in the one experiment they describe in English-German translation.

The *Champollion* system of Smadja et al. (1996) is also asymmetrical, but it addresses the harder problem of flexible collocations as well as fixed phrases. They report accuracies of 65–78% in four different experiments on the French-English Canadian Hansard parliamentary proceedings, for the equivalent of our "good input". A meaningful sense of coverage is difficult to establish, but they note that their test data includes only source language collocations with at least 10 occurrences in the corpus. In comparison, our accuracy at 99% coverage on good input was 84–92% (depending on whether we look at just the "hard" data or all the data), with 62% of our source language phrases only occurring once in the corpus.

The rest of the work on phrase translation we have found is all of the symmetrical sort. In one sense this makes the task more difficult, since source language phrases have to be discovered as well as target language phrases. On the other hand, coverage claims are often harder to evaluate since, lacking annotated test data, there is no way to tell how many more phrases a better phrase finder would have discovered that would be mistranslated by the translation finder.

Kupiec (1993) seems to have carried out the first experiments in this tradition, describing a method for finding noun phrase translations in the Canadian Hansards. Kupiec reports both accuracy and coverage: 90% accuracy, but at only 2% coverage.

Yamamoto et al. (2001) report on a symmetrical method in which the units discovered are not intended to correspond to standard syntactic phrases, which means they could not serve one of our goals, that of adding well-formed phrases to the target language lexicon. They report 83% accuracy and 60% coverage on a Japanese-English task, where coverage is ambitiously defined with respect to the entire test corpus. Their units include single words in addition to longer segments, however, and they also state that the coverage is measured automatically on an unseen corpus, which suggests that they have not verfied that their "coverage" represents correct coverage.

Wu's (1995) method, like Yamamoto et al.'s produces translation units that do not always correspond to standard syntactic phrases. He reports accuracy of 81.5% for English-Chinese, but this is for translation pairs that have survived several heuristic filters, so coverage is once again problematical.

Finally, Melamed's (1997) work on finding noncompositional compounds in parallel data focuses more on phrase finding than phrase translation. For translation finding, he simply uses previous statistical translation methods. Like Yamamoto et al. and Wu, his multiword compounds are not phrases in the traditional sense, so they would not help with our parsing problem. Finally, his goal is not to produce a phrasal lexicon, but simply to add phrase-like units to a statistical translation model, and his evaluation is in terms of improved overall performance of that model, rather than accuracy and coverage of a list of translation terms.

None of this work resembles our approach in much detail. Dagan and Church's translation-proposing method somewhat resembles a crude version of our Model 1, and Kupiec's method is somewhat like our Model 3 (replacing log-likihood-ratio scores with joint probabilities and Viterbi re-estimation with EM); otherwise, all the methods are quite different. Comparing performance is virtually impossible, since all the tasks are different and comparing coverage is extremely problematic. Nevertheless, our high accuracies at very high coverage for named-entity phrases seems to compare favorably with any of this work.

6 Conclusions

We have presented a new approach for automatically learning phrase translations from parallel corpora. Although we have tested it only on named-entity phrases, the method itself is quite general and could be applied to a wide variety of phrase translation tasks with minimal modifications. Our analysis of the "hard" subset of our data suggests that it would perform well on other tasks. The only significant change that would be need would be to generalize (or eliminate) the capitalization scores to condition on the capitalization pattern of the source language phrase, which is currently not done, since all the source language test phrases in our task had similar capitalization. With that generalization, the only obvious restriction on the applicability of the approach is that it requires the target language translations of source language phrases to be contiguous.

We plan to continue working on improving the models, including designing a proper generative probabilistic model using the features that have proved successful in the current algorithm. Finally, we plan to address the selection of source language phrases, both to correct the tokenization errors we currently make, and to extend the applicability of the method beyond named entities.

References

Y. Al-Onaizan and K. Knight. 2002. Translating named entities using monolingual and blingual resources. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, pp. 400–408.

N. Chinchor and E. Marsh. 1997. MUC-7 named entity task definition. In *Proceedings of the 7th Message Understanding Conference*, http://www.itl.nist.gov/iaui/894.02/related_projects/muc.

I. Dagan and K. Church. 1997. *Termight*: coordinating humans and machines in bilingual terminology acquisition. *Machine Translation*, 12:89–107.

T. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

J. Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, pp. 17–22.

I. D. Melamed. 1997. Automatic discovery of non-compositional compounds in parallel data. In *Proceedings of the 2nd Conference on Enpirical Methods in Natural Language Processing (EMNLP '97)*, Providence, RI.

I. D. Melamed. 2000. Models of Translational Equivalence. *Computational Linguistics*, 26(2):221–249.

F. Smadja, K. R. McKeown, and V. Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: a statistical approach. *Computational Linguistics*, 22(1):1–38.

D. Wu. 1995. Grammarless extraction of phrasal translation examples from parallel texts. in *Proceedings of TMI-95, Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, Leuven, Belgium, Vol. 2, pp. 354–372.

K. Yamamoto, Y. Matsumoto, and M. Kitamura. 2001. A comparative study on translational units for bilingual lexicon extraction. In *Proceedings of the Workshop on Data-Driven Machine Translation*, 39th Annual Meeting of the Association for Computational Linguistics, Toulouse, France, pp. 87–94.