Translating Collocations for Use in Bilingual Lexicons

Frank Smadja and Kathleen McKeown

Computer Science Department
Columbia University
New York, NY 10027
(smadja/kathy)@cs.columbia.edu

ABSTRACT

Collocations are notoriously difficult for non-native speakers to translate, primarily because they are opaque and can not be translated on a word by word basis. We describe a program named Champollion which, given a pair of parallel corpora in two different languages, automatically produces translations of an input list of collocations. Our goal is to provide a tool to compile bilingual lexical information above the word level in multiple languages and domains. The algorithm we use is based on statistical methods and produces p word translations of n word collocations in which n and p need not be the same; the collocations can be either flexible or fixed compounds. For example, Champollion translates "to make a decision," "employment equity," and "stock market," respectively into: "prendre une decision," "équité en matière d'emploi," and "bourse." Testing and evaluation of Champollion on one year's worth of the Hansards corpus yielded 300 collocations and their translations, evaluated at 77% accuracy. In this paper, we describe the statistical measures used, the algorithm, and the implementation of Champollion, presenting our results and evaluation.

1. Introduction

Hieroglyphics remained undeciphered for centuries until the discovery of the Rosetta Stone in the beginning of the 19th century in Rosetta, Egypt. The Rosetta Stone is a tablet of black basalt containing parallel inscriptions in three different writings; one in greek, and the two others in two different forms of ancient Egyptian writings (demotic and hieroglyphics). Jean-Francois Champollion, a linguist and egyptologist, made the assumption that these inscriptions were parallel and managed after several years of research to decipher the hyerogliphic inscriptions. He used his work on the Rosetta Stone as a basis from which to produce the first comprehensive hyeroglyphics dictionary.

In this paper, we describe a modern version of a similar approach: given a large corpus in two languages, our program, Champollion, produces translations of common word pairs and phrases which can form the basis for a bilingual lexicon. Our focus is on the use of statistical methods for the translation of multi-word expressions, such as collocations, which cannot consistently be translated on a word by word basis. Bilingual collocation dictionaries are currently unavailable even in languages such as French and English despite the fact that collocations have been recognized as one of the main obstacles to second language acquisition [15].

We developed a program, Champollion, which translates collocations using an aligned parallel bilingual corpus, or *database* corpus, as a reference. It represents Champollion's knowledge of both languages. For a given source language collocation, Champollion uses statistical methods to incrementally construct the collocation translation, adding one word at a time. Champollion first identifies individual words in the target language which are highly correlated with the source collocation. Then, it identifies any pairs in this set of individual words which are highly correlated with the source collocation. Similarly, triplets are produced by adding a word to a pair if it is highly correlated, and so forth until no higher combination of words is found. Champollion selects as the target collocation the group of words with highest cardinality and correlation factor. Finally, it orders the words of the target collocation by examining samples in the corpus. If word order is variable in the target collocation, Champollion labels it as *flexible* (as in *to take steps to* which can appear as: *took steps to, steps were taken to, etc.*).

To evaluate Champollion, we used a collocation compiler, Xtract[12], to automatically produce several lists of source (English) collocations. These source collocations contain both flexible word pairs which can be separated by an arbitrary number of words, and fixed constituents, such as compound noun phrases. We then ran Champollion on separate corpora, each consisting of one year's worth of data extracted from the Hansards Corpus. We asked several humans who are conversant in both French and English to judge the results. Accuracy was rated at 77% for one test set and 61% for the second set. In our discussion of results, we show how problems for the second test set can be alleviated.

In the following sections, we first describe the algorithm and statistics used in Champollion, we then present our evaluation and results, and finally, we move to a discussion of related work and our conclusions.

2. Champollion: Algorithm and Statistics

Champollion's algorithm relies on the following two assumption:

- If two groups of words are translations of one another, then the number of paired sentences in which they appear in the database corpus is greater than expected by chance. In other words, the two groups of words are correlated.
- If a set of words is correlated with the source collocation, its subsets will also be correlated with the source collocation.

The first assumption allows us to use a correlation measure as a basis for producing translations, and the second assumption allows us to reduce our search from exponential time to constant time (on the size of the corpus) using an iterative algorithm. In this section, we first describe prerequisites necessary before running Champollion, we then describe the correlation statistics, and finally we describe the algorithm and its implementation.

2.1. Preprocessing.

There are two steps that must be carried out before running Champollion. The database corpus must be aligned sentence wise and a list of collocations to be translated must be provided in the source language.

Aligning the database corpus Champollion requires that the data base corpus be aligned so that sentences that are translations of one another are co-indexed. Most bilingual corpora are given as two separate (sets of) files. The problem of identifying which sentences in one language correspond to which sentences in the other is complicated by the fact that sentence order may be reversed or several sentences may translate a single sentence. Sentence alignment programs (*i.e.,* [10], [2], [11], [1], [4]) insert identifiers before each sentence in the source and the target text so that translations are given the same identifier. For Champollion, we used corpora that had been aligned by Church's sentence alignment program [10] as our input data.[1]

Providing Champollion with a list of source collocations A list of source collocations can be compiled manually by experts, but it can also be compiled automatically by tools such as Xtract [17], [12]. Xtract produces a wide range of collocations, including flexible collocations of the type "to make a decision," in which the words can be inflected, the word order might change and the number of additional words can vary. Xtract also produces compounds, such as "The Dow Jones average of 30 industrial stock," which are rigid collocations. We used Xtract to produce a list of input collocations for Champollion.

2.2. Statistics used: The Dice coefficient.

There are several ways to measure the correlation of two events. In information retrieval, measures such as the cosine measure, the Dice coefficient, and the Jaccard coefficient have been used [21], [5], while in computational linguistics mutual information of two events is most widely used (i.e., [18], [19]). For this research we use the Dice coefficient because it offers several advantages in our context.

Let x and y be two basic events in our probability space, representing the occurrence of a given word (or group of words) in the English and French corpora respectively. Let f(x) represent the frequency of occurrence of event x, i.e., the number of sentences containing x. Then p(x), the probability of event x, can be estimated by f(x) divided by the total number of sentences. Similarly, the joint probability of x and y, $p(x \wedge y)$ is the number of sentences containing x in their English version and y in their French version ($f(x \wedge y)$) divided by the total number of sentences. We can now define the Dice coefficient and the mutual information of of x and y as:

$$Dice(x,y) = A \times \frac{2 \times (f(x \wedge y))}{f(x)+f(y)}$$
$$MU(x,y) = log(\frac{f(x \wedge y)}{f(x) \times f(y)}) + B$$

In which A and B are constants related to the size of the corpus.

We found the Dice Coefficient to be better suited than the more widely used mutual information to our problem. We are looking for a clear cut test that would decide when two events are correlated. Both for mutual information and the Dice coefficient this involves comparison with a threshold that has to be determined by experimentation. While both measures are similar in that they compare the joint probability of the two events ($p(x \wedge y)$) with their independent probabilities, they have different asymptotic behaviors. For example,

- when the two events are perfectly independent, $p(x \wedge y) = p(x) \times p(y)$.
- when one event is fully determined by the other (y occurs when and only when, x occurs), $p(x \wedge y) = p(x)$.

In the first case, mutual information is equal to a constant and is thus easily testable, whereas the Dice coefficient is equal to $\frac{2 \times (f(x) \times f(y))}{f(x)+f(y)}$ and is thus a function of the individual frequencies of x and y. In this case, the test is easier to decide when using mutual information. In case two, the results are reversed; mutual information is equal to: $-log(f(x))$ and thus grows with the inverse of the individual frequency of x, whereas the Dice coefficient is equal to a constant. Not only is the test is easier to decide using the Dice Coefficient in this case, but also note that low frequency events will have higher mutual information than high frequency events, a counter-intuitive result. Since we are looking for a way to identify correlated events we must be able to easily identify the coefficient when the two events are perfectly correlated as in case two.

Another reason that mutual information is less appropriate for our task than the Dice Coefficient is that it is, by definition, symmetric, weighting equally one-one and zero-zero matches, while the Dice Coefficient gives more weight to one-one matches. One-one matches are cases where both source and target words (or word groups) appear in corresponding sentences, while in zero-zero matches, neither source nor target words (or word groups) appear.

In short, we prefer the use of the Dice coefficient because it is a better indicator of similarity. We confirmed the performance of the Dice over mutual information experimentally as well. In our tests with a small sample of collocations, the Dice Coefficient corrected errors introduced by mutual information and never contradicted mutual information when it was correct [20].

2.3. Description of the algorithm.

For a given source collocation, Champollion produces the target collocation by first computing the set of single words that are highly correlated with the source collocation and then searching for any combination of words in that set with a high correlation with the source. In order to avoid computing and testing every possible combination which would yield a search space equal to the powerset of the set of highly correlated individual words, Champollion iteratively searches the set of combinations containing n words by adding one word from the original set to each combination of (n -1) word that has been identified as highly correlated to the source collocation. At each stage, Champollion throws out any combination with a low correlation, thereby avoiding examining any supersets of that combination in a later stage. The algorithm can be described more formally as follows:

Notation: L1 and L2 are the two languages used, and the following symbols are used:

- S: source collocation in L1
- T: target collocation in L2

[1] We are thankful to Ken Church and the Bell Laboratories for providing us with a prealigned Hansards corpus.

- WS: list of L2 words correlated with S
- P(WS): powerset of WS
- n: number of elements of P(WS)
- CC: list of candidate target L2 collocations
- P(i, WS): subset of P(WS) containing all the i-tuples
- CT: correlation threshold fixed by experimentation.

Step 1: Initialization of the work space. Collect all the words in L2 that are correlated with S, producing WS. At this point, the search space is P(WS); i.e., T is an element of P(WS). Champollion searches this space in Step 2 in an iterative manner by looking at groups of words of increasing cardinality.

Step 2: Main iteration.

$\forall$ i in {1, 2, 3, ..., n}

1. Construct P(i, WS).
   P(i, WS) is constructed by considering all the i-tuples from P(WS) that are supersets of elements of P(i-1, WS). We define P(0, WS) as null.
2. Compute correlation scores for all elements of P(i, WS). Eliminate from P(i, WS) all elements whose scores are below CT.
3. If P(i, WS) is empty exit the iteration loop.
4. Add the element of P(i,WS) with highest score to CC.
5. Increment i and go back to beginning of the iteration loop item 1.

Step 3: Determination of the best translation. Among all the elements of CC select as the target collocation T, the element with highest correlation factor. When two elements of CC have the same correlation factor then we select the one containing the largest number of words.

Step 4: Determination of word ordering. Once the translation has been selected, Champollion examines all the sentences containing the selected translation in order to determine the type of the collocation, i.e., if the collocation is flexible (i.e., word order is not fixed) or if the collocation is rigid. This is done by looking at all the sentences containing the target collocation and determining if the words are used in the same order in the majority of the cases and at the same distance from one another. In cases when the collocation is rigid, then the word order is also produced. Note that although this is done as a post processing stage, it does not require rereading the corpus since the information needed has already been precomputed.

Example output of Champollion is given in Table 1. Flexible collocations are shown with a "..." indicating where additional, variable words could appear. These examples show cases where a two word collocation is translated as one word (e.g., "health insurance"), a two word collocation is translated as three words (e.g., "employment equity"), and how words can be inverted in the translation (e.g., "advance notice").

3. Evaluation

We are carrying out three tests with Champollion with two data base corpora and three sets of source collocations. The first data base corpus (DB1) consist of 8 months of Hansards aligned data taken from 1986 and the second data base corpus consists of all of the 1986 and 1987 transcripts of the Canadian Parliament. The first set of source collocations (C1) are 300 collocations identified by Xtract on all data from 1986, the second set (C2) is a set of 300 collocations identified by Xtract on all data from 1987, and the third set of collocations (C3) consists of 300 collocations identified by Xtract on all data from 1988. We used DB1 with both C1 (experiment 1) and C2 (experiment 2) and are currently using DB2 on C3 (experiment 3). Results from the third experiment were not yet available at time of publication.

Experiment	OK	X	W	Overall
C1/DB1	70	11	19	77
C2/DB1	58	11	31	61

Table 2: Evaluation results for Champollion.

We asked three bilingual speakers to evaluate the results for the different experiments and the results are shown in Table 2. The second column gives the percentage of correct translations, the third column gives the percentage of Xtract errors, the fourth column gives the percentage of Champollion's errors, and the last column gives the percentage of Champollion's correct translation if the input is filtered of errors introduced by Xtract. Averages of the three evaluators' scores are shown, but we noted that scores of individual evaluators were within 1-2% of each other; thus, there was high agreement between judges. The best results are obtained when the data base corpus is also used as a training corpus for Xtract; ignoring Xtract errors the evaluation is as high as 77%. The second experiment produces low results as many input collocations did not appear often enough in the database corpus. We hope to show that we can compensate for this by increasing the corpus size in the third experiment.

One class of Champollion's errors arises because it does not translate closed class words such as prepositions. Since the frequency of prepositions is so high in comparison to open class words, including them in the translations throws off the correlations measures. Translations that should have included prepositions were judged inaccurate by our evaluators and this accounted for approximately 5% of the errors. This is an obvious place to begin improving the accuracy of Champollion.

4. Related Work.

The recent availability of large amounts of bilingual data has attracted interest in several areas, including sentence alignment [10], [2], [11], [1], [4], word alignment [6], alignment of groups of words [3], [7], and statistical translation [8]. Of these, aligning groups of words is most similar to the work reported here, although we consider a greater variety of groups. Note that additional research using bilingual corpora is less related to ours, addressing, for example, word sense disambiguation in the source language by examining different translations in the target [9], [8].

One line of research uses statistical techniques only for machine translation [8]. Brown *et. al.* use a stochastic language model based on the techniques used in speech recognition [19], combined with translation probabilities compiled on the aligned corpus in order to do sentence translation. The project produces high quality

English	French Equivalent
advance notice	prévenu avance
additional cost	coûts supplémentaires
apartheid ... South Africa	apartheid ... afrique sud
affirmative action	action positive
collective agreement	convention collective
free trade	libre-échange
freer trade	libéralisation ... échanges
head office	siège social
health insurance	assurance-maladie
employment equity	équité ... mati'ere ... emploi
make a decision	prendre ... décisions
to take steps	prendre ... mesures
to demonstrate support	prouver .. adhésion

Table 1: Some Translations produced by Champollion.

translations for shorter sentences (see Berger *et. al.*, this volume, for information on most recent results) using little linguistic and no semantic information. While they also align groups of words across languages in the process of translation, they are careful to point out that such groups may or may not occur at constituent breaks in the sentence. In contrast, our work aims at identifying syntactically and semantically meaningful units, which may either be constituents or flexible word pairs separated by intervening words, and provides the translation of these units for use in a variety of bilingual applications. Thus, the goals of our research are somewhat different.

Kupiec [3] describes a technique for finding noun phrase correspondences in bilingual corpora. First, (as for Champollion), the bilingual corpus must be aligned sentence-wise. Then, each corpus is run through a part of speech tagger and noun phrase recognizer separately. Finally, noun phrases are mapped to each other using an iterative reestimation algorithm. In addition to the limitations indicated in [3], it only handles NPs, whereas collocations have been shown to include parts of NPs, categories other than NPs (e.g., verb phrases), as well as flexible phrases that do not fall into a single category but involve words separated by an arbitrary number of other words, such as *"to take .. steps," "to demonstrate ... support,"* etc. In this work as in earlier work [7], we address this full range of collocations.

5. Conclusion

We have presented a method for translating collocations, implemented in Champollion. The ability to compile a set of translations for a new domain automatically will ultimately increase the portability of machine translation systems. The output of our system is a bilingual lexicon that is directly applicable to machine translation systems that use a transfer approach, since they rely on correspondences between words and phrases of the source and target languages. For interlingua systems, translating collocations can aid in augmenting the interlingua; since such phrases cannot be translated compositionally, they indicate where concepts representing such phrases must be added to the interlingua.

Since Champollion makes few assumptions about its input, it can be used for many pairs of languages with little modification. Champollion can also be applied to many domains of applications since it incorporates no assumptions about the domain. Thus, we can obtain domain specific bilingual collocation dictionaries by applying Champollion to different domain specific corpora. Since collocations and idiomatic phrases are clearly domain dependent, the facility to quickly construct the phrases used in new domains is important. A tool such as Champollion is useful for many tasks including machine (aided) translation, lexicography, language generation, and multilingual information retrieval.

6. Acknowledgements

Many thanks to Vasilis Hatzivassiloglou for technical and editorial comments. We also thank Eric Siegel for his comments on a draft of this paper. This research was partially supported by a joint grant from the Office of Naval Research and the Defense Advanced Research Projects Agency under contract N00014-89-J-1782 and by National Foundation Grant GER-90-2406.

References

1. Chen, S., "Aligning Sentences in Bilingual Corpora Using Lexical Information", *Proceedings of the 31st meeting of the ACL*, Association for Computational Linguistics, 1993, p. 9-16.
2. Church, K., "Char_align: A Program for Aligning Parallel Texts at the Character Level", *Proceedings of the 31st meeting of the ACL*, Association for Computational Linguistics, 1993, p. 1-8.
3. Kupiec, J., "An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora", *Proceedings of the 31st meeting of the ACL*, Association for Computational Linguistics, 1993, p. 17-22.
4. Simard, M., Foster, G., and Isabelle, P., "Using Cognates to Align Sentences in Bilingual Corpora", *Proceedings of the 31st meeting of the ACL*, Association for Computational Linguistics, 1993, p. 17-22.
5. Frakes, W., *Information Retrieval. Data Structures and Algorithms*, ed. W. Frakes and R. Baeza-Yates, Prentice Hall, 1992.
6. Gale, W. and Church, K., "Identifying word correspondences in parallel texts", *Darpa Speech and Natural Language Workshop*, Defense Advanced Research Projects Agency, 1991.
7. Smadja, F., "How to Compile a Bilingual Collocational Lexicon Automatically", *Proceedings of the AAAI Workshop on Statistically-Based NLP Techniques*, 1992.

8. Brown, P., Pietra, S., Pietra, V, and Mercer, R., "Word-Sense Disambiguation Using Statistical Methods", *Proceedings of the 29th meeting of the ACL*, Association for Computational Linguistics, 1991, p. 169-184.

9. Dagan, I., Itai, A., and Schwall, U., "Two Languages are more informative than one", *Proceedings of the 29th meeting of the ACL*, Association for Computational Linguistics, 1991, p. 130-137.

10. Gale, W. and Church, K., "A Program for Aligning Sentences in Bilingual Corpora.", *Proceedings of the 29th meeting of the ACL*, Association for Computational Linguistics, 1991, p. 177-184.

11. Brown, P., Lai, J. and Mercer, R., "Aligning Sentences in Parallel Corpora", *Proceedings of the 29th meeting of the ACL*, Association for Computational Linguistics, 1991, p. 169-184.

12. Smadja, F., "Retrieving collocations from text: **XTRACT**", *The Journal of Computational Linguistics*, 1993.

13. Benson, M., "Collocations and Idioms", *Dictionaries, Lexicography and Language Learning*, ed. R. Ilson, Pergamon Institute of English, 1985.

14. Benson, M., Benson, E. and Ilson, R., *The BBI Combinatory Dictionary of English: A Guide to Word Combinations*, John Benjamins, 1986.

15. Leed, R. L. and Nakhimovsky, A. D., "Lexical Functions and Language Learning ", *Slavic and East European Journal*, Vol. 23, No. 1, 1979.

16. Smadja, F., *Retrieving Collocational Knowledge from Textual Corpora. An Application: Language Generation.*, Computer Science Department, Columbia University, 1991.

17. Smadja, F. and McKeown, K., "Automatically Extracting and Representing Collocations for Language Generation", *Proceedings of the 28th annual meeting of the ACL*, Association for Computational Linguistics, 1990.

18. Church, K. and Gale, W. and Hanks, P. and Hindle, D., "Using Statistics in Lexical Analysis", *Lexical Acquisition: Using online resources to build a lexicon*, ed. Uri Žernik, Lawrence Erlbaum, 1991.

19. Bahl, L. and Brown, P. and de Souza, P. and Mercer, R., "Maximum Mutual Information of Hidden Markov Model Parameters", *Proceedings of the IEEE Acoustics, Speech and Signal Processing Society (ICASSP)*, The Institute of Electronics and Communication Engineers of Japan and The Acoustical Society of Japan, 1986, p. 49.

20. Smadja, F. and McKeown, K., "Champollion: An Automatic Tool for Developing Bilingual Lexicons," in preparation.

21. Salton, G. and McGill, M. J., *Introduction to Modern Information Retrieval*, McGraw Hill, 1983.

22. Zipf, G. K., *Human Behavior and the Principle of Least Effort*, Addison-Wesley, 1949.

23. Church, K., "Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text", *Proceedings of the Second Conference on Applied Natural Language Processing*, 1988.

24. Halliday, M.A.K., "Lexis as a Linguistic Level", *In memory of J.R. Firth*, Longmans Linguistics Library, 1966, p. 148-162.