AUTOMATIC ALIGNMENT IN PARALLEL CORPORA

Harris Papageorgiou, Lambros Cranias, Stelios Piperidis[1]

Institute for Language and Speech Processing
22, Margari Street, 115 25 Athens, Greece
Stelios.Piperidis@eurokom.ie

ABSTRACT

This paper addresses the alignment issue in the framework of exploitation of large bi-multilingual corpora for translation purposes. A generic alignment scheme is proposed that can meet varying requirements of different applications. Depending on the level at which alignment is sought, appropriate surface linguistic information is invoked coupled with information about possible unit delimiters. Each text unit (sentence, clause or phrase) is represented by the sum of its content tags. The results are then fed into a dynamic programming framework that computes the optimum alignment of units. The proposed scheme has been tested at sentence level on parallel corpora of the CELEX database. The success rate exceeded 99%. The next steps of the work concern the testing of the scheme's efficiency at lower levels endowed with necessary bilingual information about potential delimiters.

INTRODUCTION

Parallel linguistically meaningful text units are indispensable in a number of NLP and lexicographic applications and recently in the so called Example-Based Machine Translation (EBMT).

As regards EBMT, a large amount of bi-multilingual translation examples is stored in a database and input expressions are rendered in the target language by retrieving from the database that example which is most similar to the input. A task of crucial importance in this framework, is the establishment of correspondences between units of multilingual texts at sentence, phrase or even word level. The adopted criteria for ascertaining the adequacy of alignment methods are stated as follows :

- an alignment scheme must cope with the embedded extra-linguistic data (tables, anchor points, SGML markers, etc) and their possible inconsistencies.

- it should be able to process a large amount of texts in linear time and in a computationally effective way.

- in terms of performance a considerable success rate (above 99% at sentence level) must be encountered in order to construct a database with truthfully correspondent units. It is desirable that the alignment method is language-independent.

- the proposed method must be extensible to accommodate future improvements. In addition, any training or error correction mechanism should be reliable, fast and should not require vast amounts of data when switching from a pair of languages to another or dealing with different text type corpora.

Several approaches have been proposed tackling the problem at various levels. [Catizone 89] proposed linking regions of text according to the regularity of word co-occurrences across texts.

[Brown 91] described a method based on the number of words that sentences contain. Moreover, certain anchor points and paragraph markers are also considered. The method has been applied to the Hansard Corpus achieving an accuracy between 96%-97%.

[Gale 91] [Church 93] proposed a method that relies on a simple statistical model of character lengths. The model is based on the observation that longer sentences in one language tend to be translated into longer sequences in the other language while shorter ones tend to be translated into shorter ones. A probabilistic score is assigned to each pair of proposed sentence pairs, based on the ratio of lengths of the two sentences and the variance of this ratio.

[1]This research was supported by the LRE I TRANSLEARN project of the European Union

Although the apparent efficacy of the Gale-Church algorithm is undeniable and validated on different pairs of languages, it faces problems when handling complex alignments. The 2-1 alignments had five times the error rate of 1-1. The 2-2 category disclosed a 33% error rate, while the 1-0 or 0-1 alignments were totally missed.

To overcome the inherited weaknesses of the Gale-Church method, [Simard 92] proposed using cognates, which are pairs of tokens of different languages which share "obvious" phonological or orthographic and semantic properties, since these are likely to be used as mutual translations.

In this paper, an alignment scheme is proposed in order to deal with the complexity of varying requirements envisaged by different applications in a systematic way. For example, in EBMT, the requirements are strict in terms of information integrity but relaxed in terms of delay and response time. Our approach is based on several observations. First of all, we assume that establishment of correspondences between units can be applied at sentence, clause, and phrase level. Alignment at any of these levels has to invoke a different set of textual and linguistic information (acting as unit delimiters). In this paper, alignment is tackled at sentence level.

THE ALIGNMENT ALGORITHM

Content words, unlike functional ones, might be interpreted as the bearers that convey information by denoting the entities and their relationships in the world. The notion of spreading the semantic load supports the idea that every content word should be represented as the union of all the parts of speech we can assign to it [Basili 92]. The postulated assumption is that a connection between two units of text is established if, and only if, the semantic load in one unit approximates the semantic load of the other.

Based on the fact that the principal requirement in any translation exercise is meaning preservation across the languages of the translation pair, we define the semantic load of a sentence as the patterns of tags of its content words. Content words are taken to be verbs, nouns, adjectives and adverbs. The complexity of transfer in translation imposes the consideration of the number of content tags which appear in a tag pattern. By considering the total number of content tags the morphological derivation procedures observed across languages, e.g. the transfer of a verb into a verb+deverbal noun pattern, are taken into account. Morphological ambiguity problems pertaining to content words are treated by constructing ambiguity classes (acs) leading to a generalised set of content tags.

It is essential here to clarify that in this approach no disambiguation module is prerequisite. The time breakdown for morphological tagging, without a disambiguator device, is according to [Cutting 92] in the order of 1000 μseconds per token. Thus, tens of megabytes of text may then be tagged per hour and high coverage can be obtained without prohibitive effort.

Having identified the semantic load of a sentence , Multiple Linear Regression is used to build a quantitative model relating the content tags of the source language (SL) sentence to the response, which is assumed to be the sum of the counts of the corresponding content tags in the target language (TL) sentence. The regression model is fit to a set of sample data which has been manually aligned at sentence level. Since we intuitively believe that a simple summation over the SL content tag counts would be a rather good estimator of the response, we decide that the use of a linear model would be a cost-effective solution.

The linear dependency of y (the sum of the counts of the content tags in the TL sentence) upon x_i (the counts of each content tag category and of each ambiguity class over the SL sentence) can be stated as :

$$y=b_0+b_1x_1+b_2x_2+b_3x_3+...+b_nx_n+\varepsilon \quad (1)$$

where the unknown parameters $\{b_i\}$ are the regression coefficients, and ε is the error of estimation assumed to be normally distributed with zero mean and variance σ^2.

In order to deal with different taggers and alternative tagsets, other configurations of (1), merging acs appropriately, are also recommended. For example, if an acs accounts for unknown words, we can use the fact that most unknown words are nouns or proper nouns and merge this category with nouns. We can also merge acs that are represented with only a few distinct words in the training corpus. Moreover, the use of relatively few acs (associated with content words) reduces the number of parameters

to be estimated, affecting the size of the sample and the time required for training.

The method of least squares is used to estimate the regression coefficients in (1). Having estimated the b_i and σ^2, the probabilistic score assigned to the comparison of two sentences across languages is just the area under the $N(0,\sigma^2)$ p.d.f., specified by the estimation error. This probabilistic score is utilised in a Dynamic Programming (DP) framework similar to the one described in [Gale 91]. The DP algorithm is applied to aligned paragraphs and produces the optimum alignment of sentences within the paragraphs.

EVALUATION

The application on which we are developing and testing the method is implemented on the Greek-English language pair of sentences of the CELEX corpus (the computerised documentation system on European Community Law).

Training was performed on 40 Articles of the CELEX corpus accounting for 30000 words. We have tested this algorithm on a randomly selected corpus of the same text type of about 3200 sentences. Due to the sparseness of acs (associated only with content words) in our training data, we reconstruct (1) by using four variables. For inflective languages like Greek, morphological information associated to word forms plays a crucial role in assigning a single category. Moreover, by counting instances of acs in the training corpus, we observed that words that, for example, can be a noun or a verb, are (due to the lack of the second singular person in the corpus) exclusively nouns. Hence :

$$y=b_0+b_1x_1+b_2x_2+b_3x_3+b_4x_4+\varepsilon \quad (2)$$

where x_1 represents verbs, x_2 stands for nouns, unknown words, vernou (verb or noun) and nouadj (noun or adjective), x_3 adjectives and veradj (verb or adjective), x_4 adverbs and advadj (adverb or adjective).

σ^2 was estimated at 3.25 on our training sample, while the regression coefficients were:

$b_0 = 0.2848, b_1 = 1.1075, b_2 = 0.9474,$
$b_3 = 0.8584, b_4 = 0.7579$

An accuracy that approximated a 100% success rate was recorded. Results are shown in Table 1. It is remarkable that there is no need for any lexical constraints or certain anchor points to improve the performance. Additionally, the same model and parameters can be used in order to cope with the infra-sentence alignment.

In order to align all the CELEX texts, we intend to prepare the material (text handling, pos tagging in different languages pairs and different tag sets, etc.) so that we will be able to evaluate the method on a more reliable basis. We also hope to test the method's efficiency at phrase level endowed with necessary bilingual information about phrase delimiters. It will be shown there, that reusability of previous information facilitates tuning and resolving of inconsistencies between various delimiters.

category	N	correct matches
1-0 or 0-1	5	4
1-1	3178	3178
2-1 or 1-2	36	35
2-2	0	0

Table 1 : Matches in sentence pairs of the CELEX corpus

REFERENCES

[Basili 92] Basili R. Pazienza M. Velardi P. "Computational lexicons: The neat examples and the odd exemplars". Proc. of the Third Conference on Applied NLP 1992

[Brown 91] Brown P. Lai J. and Mercer R. "Aligning sentences in parallel corpora". Proc. of ACL 1991

[Catizone 89] Catizone R. Russell G. Warwick S. "Deriving translation data from bilingual texts". Proc. of the First Lexical Acquisition Workshop, Detroit 1989

[Church 93] Church K. "Char_align: A program for aligning parallel texts at character level" Proc. of ACL 93

[Cutting 92] Cutting D. Kupiec J. Pedersen J. Sibun P. "A practical part-of-speech tagger " Proc.of ACL 1992

[Gale 91] Gale W. Church K. "A program for aligning sentences in bilingual corpora". Proc. of ACL 1991

[Simard 92] Simard M. Foster G. Isabelle P. "Using cognates to align sentences in bilingual corpora" Proc. of TMI 1992