Cautious Steps towards Hybrid Connectionist Bilingual Phrase Alignment

Stefan Wermter
International Computer Science Institute
1947 Center Street
Berkeley, CA 94704-1198
USA
wermter@icsi.berkeley.edu

Joseph Chen
University of Hamburg
Computer Science Department
Vogt-Kölln-Straße 30
D-22527 Hamburg, Germany
chen@informatik.uni-hamburg.de

Abstract

[1]We examine phrase alignment in a hybrid connectionist framework. We describe the architecture and the learning algorithm of our approach. Simulations have been carried out to demonstrate the feasibility of this approach using a real-world title phrase corpus. Although the results of our approach are still at an early stage, we found that a hybrid approach to phrase alignment has the potential to provide good results using relatively little training data. While there have been a lot of statistical approaches to alignment, to the best of our knowledge this is the first hybrid connectionist approach for phrase alignment.

1 Introduction

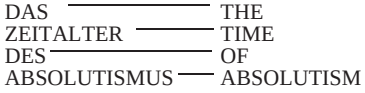
Recently, the field of computational linguistics has seen a remarkable trend for certain tasks. Rather than using traditional linguistic symbolic grammar formalisms new learning approaches have been examined, especially for restricted tasks and domains which may not need a deep structural analysis. Among these alternative approaches are statistical (Chen 96; Brown *et al.* 93; Brown *et al.* 92), connectionist (Elman 90; Chalmers 90; John & McClelland 90; Munro *et al.* 91) and hybrid approaches (Wermter & Weber 97; Wermter 95). These approaches have been used for learning restricted tasks like word sense disambiguation, syntactic tagging, language modeling, etc. We believe that these restricted tasks provide an appropriate framework for the applicability of new hybrid connectionist learning approaches in natural language processing.

In this paper we describe initial experiments using a hybrid statistical/symbolic/connectionist approach for the alignment and collocation in bilingual corpora. Previous approaches for alignment are mostly statistical approaches (Chen 96; Gale & Church 93; Wu 94; Fung & McKeown 94; Kay & Röscheisen 93; Xu & Tan 96). These approaches collect the statistical characteristics from large bilingual corpora and construct a model for the context without making much effort to analyze the sentences syntactically or semantically. That is, purely statistical approaches "learn" associations between lexical items (words) rather than using syntax or semantics. Therefore, purely statistical approaches need extremely large bilingual corpora so that they can estimate the probability that a certain lexical item (word) is associated with a certain lexical item in a different language. However, such large bilingual corpora of gigabyte size or more are hard to collect. Therefore, the question arises whether there are other learning approaches which use more input knowledge but much less training data.

In this paper, we use a hybrid statistical/symbolic/connectionist framework to attack the alignment problem. Our hybrid approach uses symbolically interpretable syntactic and morphological input and output representations for words in order to support a reduction of the training data compared with the large number of lexical words in other statistical approaches. Furthermore, our hybrid approach uses statistical techniques for data compression only as well as connectionist techniques for the main learning part. The learnability and robustness of a connectionist framework appears to be a good candidate for a robust alignment task. On the other hand, the readily available symbolic information for the language representations can still be used. This motivates our hybrid approach to the alignment problem.

2 Alignment for Language Translation

Alignment is the problem to find corresponding parts in sentences or phrases from two languages. For example, assume the following two phrases have to be aligned.



In this example the corresponding words have been aligned by lines to demonstrate their correspondence. In general, it is possible to translate such phrases differently but we assume that a reasonable alignment is given and our goal is to try to learn and generalize this alignment as well as possible.

There are many other phenomena which have to be considered for language alignment and language translation. For instance, one word may not have a corresponding word in another language and therefore it has to be circumscribed. For example, from German

[1]Appeared in R. Mitkov, N. Nicolov and N. Nikolov Ed. *Proceedings International Conference Recent Advances in Natural Language Processing* pp. 364-368, Tzigov Chark, Bulgaria, 1997

to Chinese complicated subordinate clauses are usually avoided and translated as several simple sentences. The single compound noun in German has to be translated into several English words, possibly in a phrasal form.

In other cases one word can be translated as many different words, depending on the context. Furthermore, morphological knowledge may differ a lot between two different languages. For example, the noun gender, verb conjugation and noun declension are not uniformly present in all languages. Functional words such as gender or number markers may have to be introduced in such cases. Some categories may be missing in one language while they remain very important in the other language. An example are the quantitatives in Chinese/Japanese and articles in English/German/French.

Our title corpus contains several thousand book titles extracted from the bibliographic entries in a library (Wermter 95). We randomly chose 100 titles from this corpus and translated them to English. The English translations were double-checked and controlled by several native American speakers. Then the German words of the titles were aligned with their corresponding English translations. For example, in the following title phrase and its translation we also give the numbers of the corresponding word positions.

DAS ZEITALTER DES ABSOLUTISMUS
THE TIME OF ABSOLUTISM
Alignment: $((1 \Leftrightarrow 1)(2 \Leftrightarrow 2)(3 \Leftrightarrow 3)(4 \Leftrightarrow 4))$

Thus (1⇔1) means we align DAS with THE and so on. In fact we could say that ((1 2 3 4 ⇔ 1 2 3 4)) and ((1 2 ⇔ 1 2) (3 4 ⇔ 3 4) are also valid alignments. However, our guideline is to provide the most detailed correspondences. Therefore we divide into as many corresponding parts as possible.

Most titles are noun phrases and prepositional phrases. Their length is from one word to thirteen words. The German titles often contain long compound nouns which have to be translated into several English words or phrases. Although many of the phrases contain word-by-word correspondences there are also translations which do not contain easy correspondences.

3 A Hybrid Architecture

3.1 Overview of the Architecture

Our initial architecture is motivated by our goal of learning robust alignment. Therefore, we want to exploit the learning capabilities of connectionist networks as well as the advantageous interpretation capabilities of symbolic input/output representations. For reducing the size of the input/output representations we want to take advantage of the compression capabilities of statistical techniques. Our proposal for a symbolic/statistical/connectionist aligner is illustrated in Figure 1. The architecture is modular so that the task

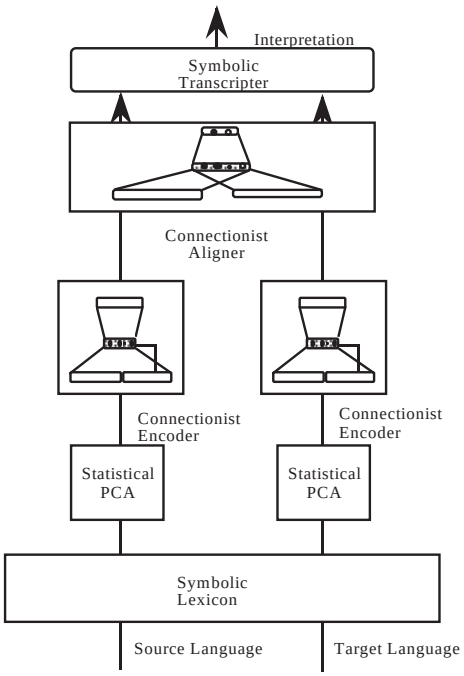


Figure 1: Modular design of a bilingual aligner

performed by a specific module can be replaced by another module as long as the input-output characteristics are preserved. This allows for easy changes in the overall hybrid architecture.

3.2 Symbolic Lexicon Module

In the title corpus a German word was represented as a 22-dimensional activation vector. Each unit in the activation vector represents a linguistic feature. The features in the German lexicon are: number, noun, proper noun, adjective, preposition, determinate article, adverb, verb, indeterminate article, pronoun, conjunctive, time modifier, masculine, feminine, neutral, plural, dative, accusative, genitive, transitive, intransitive, ordinal. These features are based on categories from the Langenscheidt dictionary.

An English word was encoded as a 16-dimensional activation vector. The features in the English lexicon are: number, noun, proper noun, adjective, preposition, article, adverb, verb, pronoun, conjunctive, time modifier, plural, genitive, intransitive, transitive, ordinal.

The English and German representations differ slightly due to differences with respect to declensions and conjugations.

The vocabulary of the German phrases contains 294 different words, and that of the English phrases 332 words. The difference is due mostly to compound nouns which have to be translated as longer phrases in English.

3.3 Statistical Preprocessing: Principal Component Analysis

Recurrent neural networks can have relatively expensive numerical computations due to their recurrent connections. So it is an important issue to keep the size of a network small in order to support fast training. One possibility to compress the local input representation of a single word is to apply principal component analysis (PCA) (Gonzalez & Wintz 77) to the input data. Then most important leading components of the transformed vector are representative for the word representation (in our experiments described below we used the first five components).

The use of the PCA can be justified by the following observation: For the application of natural language processing, a word could have many linguistic features but it does not actually have most of them. In other words, the features are sparse and only a few specific feature combinations are characteristic for a particular word. For example, a word is very unlikely to have the features "verb" and "animate" at the same time, while it is much more likely for the features "verb" and "past-tense" or "noun" and "animate". Furthermore, the distribution of vocabulary is very different from the uniform distribution. PCA takes into account the occurrence frequency of the actual vocabulary. The eigenbase of the PCA can be used to reconstruct the decoded item later to its original dimension.

3.4 Encoding Symbolic Sequences in Connectionist Networks

3.4.1 Context Representation Modules

The central modules in this architecture are the artificial neural networks (ANNs) that learn the crucial encoding tasks in alignment. We examine two types of networks for sequential processing within the overall hybrid architecture: a recurrent auto-associative memory (RAAM) (Pollack 90) and a simple recurrent network (SRN) (Elman 90).

In the RAAM version of the architecture, there are two ANN encoders and decoders, one pair for each language (see also Figure 1 and Figure 2). Each encoder and decoder is trained to auto-associate the words in each phrase/utterance. We coarsely divide the word representation into two locally encoded feature parts: a major part of speech code and a minor morphologic/functional code. Both are referred to as generalized part of speech representation (abbreviated as GPOS). The two ANN are working independently and can develop different compressions of phrases or sentences according to the specific language. The RAAM architecture is illustrated in Figure 2.

In the second version of our architecture we use simple recurrent networks (Elman 90) as an alternative for encoding and decoding the contextual information in a phrase (see Figure 3). The contextual information gracefully degrades to the left of the current word, which makes this model more plausible than a fixed

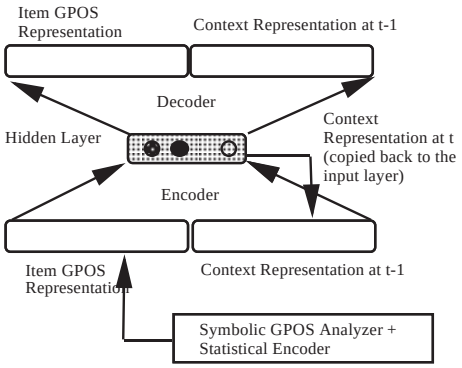


Figure 2: RAAM encoder architecture

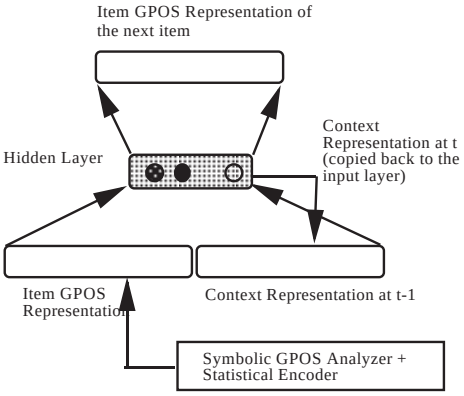


Figure 3: SRN encoder architecture

n-gram model. More specifically, the activation in the hidden layer encodes the left context from the beginning of the phrase to the current word. While the RAAM networks are used as autoassociators of a sequence of words these SRN are used to predict the representation of the next word in a sequence. The last word in the phrase should predict a particular symbol for the end of the phrase which is represented by the average activation of each unit. The context layer is reset between two different phrases.

In our implementation, the neural networks are trained with the conjugate gradient method (Press *et al.* 92).

3.5 FF-Networks for Aligning RAAM and SRN Representations

So far we have described the overall hybrid architecture, a statistical preprocessor for dimension reduction and two alternative versions to encode partial sequences of words as RAAMs and SRNs. We will now focus on the issue of learning the alignment itself. The alignment is considered to be a binary relation which is bidirectional. We use a feedforward network as illustrated in Figure 4. This network is motivated by the observation that the available information for an effective alignment is the representation of a word in

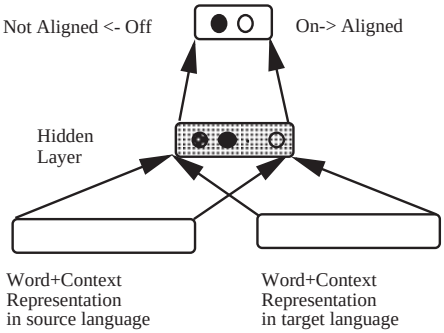


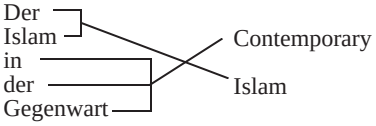
Figure 4: Feedforward network as an aligner

its context.

The word representation and the context representation, which is the activation in the hidden layer of either SRN or RAAM encoder, are concatenated as the input vector for the feedforward network. In the SRN implementation, the context representation is the hidden activation at the current word in a phrase, and the word representation is the feature representation. In the RAAM implementation, the context representation is the final hidden activation. The word representation, however, is the hidden activation at the current word. This schema is chosen because the RAAM can decode the word itself given a hidden activation. The concatenation is denoted as "Word+Context Representation" in Figure 4 .

4 Experiments with Various Hybrid Architectures

Our measurement of correctness is based on the number of individual alignments. The individual alignment is a set of word pairs and can be denoted by a pair of integers. Each integer represents the index number of the word in the German and English phrase, respectively. Consider the following example:

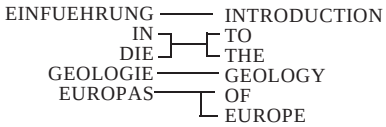


There are 5 x 2 = 10 individual alignments, each of which can be either positive or negative. In the example above, the individual alignment (Der $\Leftrightarrow$ Islam) (or (1 $\Leftrightarrow$ 2)) and (Islam $\Leftrightarrow$ Islam) (or (2 $\Leftrightarrow$ 2)) are positive; the individual alignment (Der $\Leftrightarrow$ Contemporary) (or (1 $\Leftrightarrow$ 1)) is negative. If an individual alignment from the output of the network is positive (or negative) as indicated by its label, this outcome is counted as correct, otherwise incorrect.

The correctness is calculated by dividing the number of correct individual alignments by the total number of individual alignments. A pair is considered as "positive" if and only if the activation of the aligned-unit is higher than that of the not-aligned-unit (see Figure 4). The ratio between the negative alignment and positive alignment pairs depends on the length of the phrase/sentence and the complexity of the language. We found that this ratio in the title phrase corpus is about 4. We correct this unbalance by using a normalization mechanism in preparing the training set. We modified the error function so that it can scale the error contributed by the positive and the negative alignments to the same degree.

There are 80 phrases in the training set and 20 phrases in the test set. The number of hidden units is 24 in both SRN context encoding and feed forward alignment network. For the training set, the alignment accuracy rate is 99.95% (2424/2425 words). The test set contains 20 phrases that were not in the training set. The overall alignment accuracy rate is 82.5% (368/446 words).

A typical example of the performance of the SRN model is:



The network geometry for the RAAM is the same as in the SRN configuration. The alignment accuracy for the training set is 100% (2439/2439). On the test set, an accuracy of 84.7% (382/451) is achieved.

The networks learn the mapping from the training set. If the test set contains very different alignments the networks will not generalize well, but they will generalize well for alignments for which similar patterns have been seen during training.

However, at the current stage, our approach can learn and generalize the alignment of smaller phrase groups quite reliably already.

5 Discussion and Conclusion

We have proposed an hybrid approach for alignment. While there have been a lot of statistical approaches to alignment, to the best of our knowledge this is the first hybrid connectionist approach for phrase alignment. This research is still at an early stage compared to many traditional alignment approaches which use mainly statistical techniques. However, the purely statistical approaches suffer from the large quantities of training material. In our approach, we try to address alignment using smaller corpora since gigabyte or terabyte size corpora are not readily available in all domains.

A symbolic lexicon component allows the association of category knowledge with lexical words. Since category representations of words can grow large we use a statistical preprocessing step to reduce the size of the input representations. Finally, all the actual sequence encoding and alignment is learned using various connectionist networks. In contrast to symbolically and manually encoded aligners connectionist networks allow the learning of alignment regularities from examples and therefore should be more robust. On the other hand, the use of symbolic category representa-

tions allows to test alignment strategies for small and medium corpora, which are areas for which statistical alignment on lexical word items does not perform well.

We have used and compared RAAM and SRN networks for encoding sequences of words. The autoassociation task in RAAMs and the prediction task in SRN were not particularly developed for the alignment task. Furthermore, the performance of these individual networks for autoassociation and prediction was not perfect. In general, this means that the encoding networks did not yet contain much domain or task knowledge. Given this restricted autoassociative or predictive context knowledge the performance of the alignment network is quite reasonable already.

This leads us to expect that this approach has some potential, especially since we still did not yet make use of important underlying knowledge for alignment (e.g. semantics), since the encoding networks did not yet use alignment specific knowledge, and since initial results with simple recurrent networks on a real-world corpus were quite encouraging. We expect that a suitable hybrid connectionist architecture which is able to systematically encode a natural language sequence with more semantics could also improve the performance scaling up to bigger phrases and sentences.

Acknowledgments

We would like to thank three anonymous reviewers for their comments; furthermore we would like to thank especially Zack Philips, Elizabeth Weinstein for proofreading the German-English translations of the title material. This research was funded by the German Research Association (DFG) under Grant DFG We 1468/4-1.

References

(Brown *et al.* 92) Peter F. Brown, V.J. Della Pietra, P.V. de Souza, J.C. Lai, and R.L. Mercer. Class-based n-gram models of natural language. *Computer Linguistics*, 18(4):467–479, 1992.

(Brown *et al.* 93) P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. The mathemtics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

(Chalmers 90) D.J. Chalmers. Syntactic transformations on distributed representations. *Connection Science*, 2(1&2):53–62, 1990.

(Chen 96) Stanley F. Chen. *Building Probabilistic Models for Natural Language*. Phd thesis, Harvard University, 1996.

(Elman 90) Jeffery Elman. Finding structure in time. *Cognitive Science*, 14:179–211, 1990.

(Fung & McKeown 94) Pascale Fung and K. McKeown. Aligning noisy parallel corpora across language groups: Word pair feature matching by dynamic time warping. e-print, CS Dept, Columbia University, N.Y., 1994.

(Gale & Church 93) William A. Gale and K.W. Church. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102, 1993.

(Gonzalez & Wintz 77) Rafael C. Gonzalez and P. Wintz. *Digital Image Processing*. Addison-Wesley Publishing Company, Reading, Massachusetts, 1977.

(John & McClelland 90) Mark F. St. John and J. L. McClelland. Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, 46:217–257, 1990.

(Kay & Röscheisen 93) Martin Kay and Martin Röscheisen. Text-translation alignment. *Computational Linguistics*, 19(1):121–142, 1993.

(Munro *et al.* 91) Paul Munro, C. Cosic, and M. Tabasko. A network for encoding, decoding and translating locative prepositions. *Connection Science*, 3(3):225–240, 1991.

(Pollack 90) Jordan B. Pollack. Recursive distributed representations. *Artificial Intelligence*, 46:77–105, 1990.

(Press *et al.* 92) William H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. Minimization or maximization of functions. In *Numerical Recipes in C*. Cambridge University Press, 1992.

(Wermter & Weber 97) Stefan Wermter and V. Weber. Screen: Learning a flat syntactic and semantic spoken language analysis using artificial neural networks. *Journal of Artificial Intelligence Research*, 6(1):35–85, 1997.

(Wermter 95) Stefan Wermter. *Hybrid Connectionist Natural Language Processing*. Chapman & Hall Neural Computing Series. Chapman & Hall, London, 1995.

(Wu 94) Dekai Wu. Aligning a parallel english-chinese corpus statistically with lexical criteria. Technical report, Dept. of Computer Science, University of Science & Technology, Clear Water Bay, Hong Kong, 1994.

(Xu & Tan 96) Donghua Xu and C. L. Tan. Automatic alignment of english-chinese bilingual texts of cns news. Technical report, Department of Information system and Computer Science, National Univerisity of Sngapore, Singapore 119260, 1996.