Statistical Machine Translation Using the Self-Organizing Map

V.F. López, J.M. Corchado, J.F. De Paz, S. Rodríguez, and J. Bajo

Abstract. The paper describes a contextual environment using the Self-Organizing Map, which can model a semantic agent (SOMAgent) that learns the correct meaning of a word used in context in order to deal with specific phenomena such as ambiguity, and to generate more precise alignments that can improve the first choice of the Statistical Machine Translation system giving linguistic knowledge.

1 Introduction

For more than half a century, various aspects of translation have been studied and considered in order to develop Machine Translation (MT). However, it is well-known that MT is a very difficult task. The more general the domain or complex the style of the text, the more difficult it is to achieve a high quality translation. Today there is a wave of optimism that is spreading throughout the MT research community, one that has been caused by the revival of statistical approaches to MT. Very specifically, we refer to the birth of Statistical Machine Translation (SMT). In contrast to previous approaches based on linguistic knowledge representation, SMT is based on large amounts of human-translated example sentences (parallel corpora) from which it is possible to estimate a set of statistical models describing the translation process [9].

The incorporation of syntactic information in SMT is a current research topic. It is based on both syntax and on hierarchy of phrases. To this end, in [9, 19] there appears the need to introduce alternative techniques to include information on morphology derivation and verb group information into word alignment algorithms.

In this paper, we study improvements in translation quality that can be achieved by using the open-source Syntax Augmented Machine Translation (SAMT). By preprocessing with a multi-agent system, we experimented with different degrees of linguistic analysis from the lexical level to a syntactic or semantic level in order to generate a more precise alignment. We developed a contextual environment using

V.F. López, J.M. Corchado, J.F. De Paz, S. Rodríguez, and J. Bajo
Dept. Informática y Automática. University of Salamanca,
Plaza de la Merced S/N, 37008. Salamanca
e-mail: vivian@usal.es

A.P. de Leon F. de Carvalho et al. (Eds.): Distrib. Computing & Artif. Intell., AISC 79, pp. 131–138.
springerlink.com © Springer-Verlag Berlin Heidelberg 2010

the Self-Organizing Map where we model a semantic agent (SOMAgent) that learns the correct meaning of a word used in a particular context in order to deal with specific phenomena such as ambiguity and to generate more precise alignments that can improve the first choice of the SMT system.

The Machine Translation and the Statistical Approach are further described in Section 2. The SAMT is presented in Section 3. The Word Alignment with SOMAgent and our system is further described in Section 4 and the conclusions are briefly outlined in Section 5.

2 Machine Translation and the Statistical Approach

SMT as a research area started in the late 1980s with the Candide Project at IBM, which included the classic IBM word-based model. Their estimation of a parallel corpus can be found in [1]. When IBM researchers presented the statistical approach to MT, the interest among both natural language and speech processing research communities increased. The IBM model included the possibility of working towards a level of phrases. The evolution from word-based models to phrase-based models is described in [10] and Moses MT ($http://www.statmt.org/moses/$). Marcu [14] introduced a joint-probability model for phrase translation. As a result, most competitive SMT systems, such as the CMU, IBM, ISI, and Google systems, to name just a few, use phrase translation. Phrase-based systems came out ahead of the participation list at a recent international MT competition (DARPA TIDES Machine Translation Evaluation 2003-2006 on Chinese-English and Arabic-English). They also appear the SMT model based on tuple N-grams [15], or Ngram-based SMT. This approach is an evolution of a previous Finite-State Transducer implementation of X-grams [2], which adapted speech recognition tools for speech-oriented MT. The result is a competitive SMT model whose basic unit is the tuple, composed by one or more words of the language source and for one or more words of the target language.

In the last year, many efforts have been devoted to building syntax-based models that use either real syntax trees generated by syntactic parsers, or tree transfer methods motivated by syntactic reordering patterns. This statistical approach had considerable success. Several other strategies have been followed, including systems based on syntax [16], and those based on the hierarchy of phrases [5].

3 Syntax Augmented Machine Translation

Defined in [22] as a specific parameterization of the probabilistic synchronous context-free grammar (PSCFG) approach to MT. It takes advantage of nonterminal symbols, as in monolingual parsing, to generalize beyond purely lexical translation. [6] extends SAMT to include nonterminal symbols from target language phrase structure parse trees. Each target sentence in the training corpus is parsed with a stochastic parser [4] to produce constituent labels for target spans. PSCFG are

defined by a source vocabulary T_s, a target vocabulary T_t, and a shared non-terminal set N, and induce rules of the form

$$X =< \gamma, \alpha, \iota, \psi > \tag{1}$$

Where $X \in N$ is a nonterminal (initial rule), $\gamma \in (NUT_s)^*$ is a sequence of nonterminals and source terminals, $\alpha \in (NUT_t)^*$ is a sequence of nonterminals and target terminals, ι is a one to one mapping from nonterminal tokens in γ to nonterminal tokens in α, and ψ is a non negative weight assigned to the rule.

PSCFG models define weighted transduction rules that are automatically learned from parallel training data. As in monolingual parsing, such rules make use of nonterminal categories to generalize beyond the lexical level. These rules seem considerably more complex than weighted word-to-word rules [1], or phrase-to-phrase rules [10] but can be viewed as natural extensions to these well established approaches. In [6] it is pointed out a procedure to learn PSCFG rules from word-aligned parallel corpora, using the phrase-pairs as a lexical basis for the grammar.

The translation quality is represented by a set of the functions for every rule, that are trained via Minimum Error Rate (MER) [17] to maximize translation quality according to a user specified automatic translation metric, like BLUE Papineni et al. [18] or NIST [8]. The weights of the functions are computed on the basis of the maximization of the BLUE measure.

4 Word Alignment

In this application it is intended to demonstrate that Kohonen Maps [12][11] can be applied to introducing linguistic information, other than the lexical units, to the process of building word and phrase alignments. We consider that linguistic information may be helpful to built better translation models. The alignment model as part of a whole translation scheme can also be defined as an independent Natural Language Processing task. In fact, most of current new generation translation models treat word alignment as an independent result from the translation model. In [19] the task of automatic word alignment focuses on detecting, given a parallel corpus, which tokens or sets of tokens from each language are connected together in a given translation context, revealing thus the relationship between these bilingual units.

4.1 The Word Alignment with SOMAgent

Our approach exploits the possibility of working with alignments at different levels of granularity, from the lexical to the semantic level, as suggests [19]. Therefore, assuming we are able to extract a set of tuples from a given parallel text, we can use a multi-agent system (SOMAgent) [13] to estimate the bilingual model and, to perform a corpus preprocessing, for SMT in a prototype of an Automatic German-Spanish Translator.

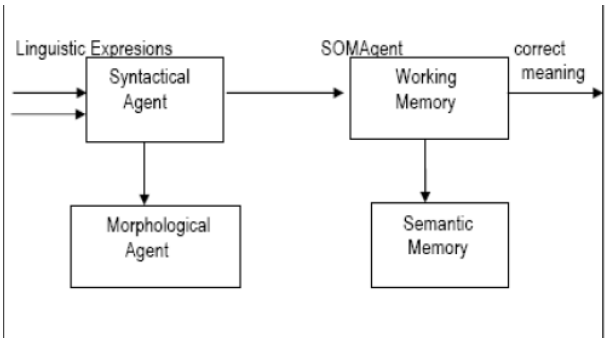

Fig. 1 Architecture of multi-agent system

The overall architecture of multi-agent system is presented in Figure 1. The SOMAgent receives perceptual inputs: linguistic expressions. There are potential actions: the agent can disambiguate an expression. The perceptions words are primarily stored in the working memory. The semantic memory associates contextual information and gives the correct meaning. Communication between the agents is motivated by the exchange of information related to linguistic expressions: morphological, syntactical and semantic information about the lexical items that are necessary for the resolution of specific tasks.

In order to implement this model, grammar knowledge comprises the initial tree models, which represent the structure of German sentences and the lexicalization dictionary forming the Syntactical Agent knowledge. This agent can be seen as a subsociety [20], formed by agents handling simpler task or information associated with the features (e.g. complements) used in the parsing. This subsociety can be dynamically organized according to the problem it is expected to solve: to assist in a best alignment. The Syntactical agent [13] divides the sentence into subject, verb, object and enrich tokens with features further than lexical such as part-of-speech (PoS),lemma,and chunk IOB label.In cases where syntactic-semantic analysis of the society of agents is insufficient to resolve a lexical ambiguity so that it should be solved by context reference. The network is trained using the SOMAgent with a large set of sentences that reflects every type of context in the corpus. These sentences, following the steps of the general algorithm [13], form a file of input data vectors for doing the training, creating the semantic memory (a trained network) with the semantic classes specified.

To study semantic relationships in their pure form, it is recognized that semantic value should not be inferred from any semantic pattern used for the encoding of individual words but only from the context where each word appears. In the self-organizing process, the inputs consist of sequences of three words selected from certain patterns of contexts. Such class patterns are defined off-line. With sentence patterns generated based on this contexts, sentences are created covering every possible context combination, for example: *Peter spielt Fuβball (Peter plays football), Peter spielt Karten (Peter plays cards) or Peter spielt Schach (Peter plays chess).*

Table 1 A case of word alignment possibilities on a top of lexical units (a) and linguistic data (b)

Peter	spielt	Fuβball,	Peter	spielt	Gitarre,	and	Peter	liebt	romantische		Spiel
↙	↙	↙	↙	↙	↙	↙	↙	↙	↙		↙
Peter	juega	futball,	Peter	toca	guitarra,	and	Peter	loves	romantic		play

The training phase consists of the sequential presentation of semantically correct sentences until the network converges. After training, the network becomes topologically ordered, and it can be verified what units of the map are active for each input vector and are then labeled, with the principal semantic classes, taking the best answer for conducting via automatic model clustering to reduce the ambiguity. For those cases, the SOMAgent is called to collaborate in solving the ambiguity, the agent takes as its input the results of the previous agents: the semantic agent searches for meanings associated with each word, forming key sentences with the combination of words in German which could not disambiguate, these feed the network input, which should be able to classify it within the active classes, taking the best answer as the correct meaning and the best alignment.

For example, suppose the case illustrated in table 1, for the sentence *Peter spielt Fuβball* we take, the German verb *spielen* (to play) that has two meanings represented by different Spanish verbs: either *tocar*, which appears in the context of playing musical instruments Klavier, Gitarre, Flöte, or *jugar* which appears in the context of games, Fuβball, Karten or Schach. In addition, the lexical item *spielen* is seen acting as a verb and as a noun. Considering these two words, with the same lexical realization, as a single token adds noise to the word alignment process. The Syntactical Agent represent this information, by syntactic label (by means of linguistic data views), as *spielen VBZ* and *spielen NNS* would allow us to distinguish between the two cases. For those cases where the Semantic Agent collaborates in solving the semantic ambiguity, the agent takes as its input the results of the previous agents, searches for meanings associated with each word, forming key sentences with the combination of words in German which could not disambiguate, these feed the network input, which should be able to classify it within the active semantic classes, taking the best answer as the correct meaning. For example for the sentence *Peter spielt Fuβball*, the network find the true meaning of German verb *spielt*, alignment this entry inside the active classes, in this case, the class 2 [13](to play) whose meaning is *jugar* in Spanish.

4.2 Experimental Work

We present the experimental results for Germanto to Spanish translation task, based on a set of sentences of the full DWDS corpus ($http://utils.mucattu.com/$) of the domain of news. The results were obtained using only the first 40K lines of the corpus. The statistical data set of the corpus can be seen in the table 2.

Table 2 Training set

	Spanish	German
Sentences	40 K	40 K
Words	1,31	1,47
Length average	18,10	31,11
Vocabulary	41,12	21,10

For phrase extraction we have used MOSES MT. The number of phrases of the style Moses extracted with the system based on phrases was 4,8M. The first preliminary step requires the preprocessing of the parallel data using SOMAgent, so that it is sentence aligned and tokenised. It has as aim to deal with specific phenomena such as ambiguity and to generate more precise alignments. The output of the tokenised is formed from words that are meaningful within a particular context (or domain). For dimensionality reduction it excludes words which are meaningless because they are independent of the domain and they belong to categories such as articles, prepositions, conjuntions and pronouns. This allows the network to be trained with a smaller range of errors. The training data were provided for the sentence aligned (one sentence per line), in two files, one for the German sentences, one for the Spanish sentences. A phrase-based translation models was built of the output of the multi-agent systems to extract the purely lexical phrases, which later were used to create the grammar of the SAMT. Then, running the script that forms part of the Moses MT System grow-diag-final aligned as well as was computed the word-to-word lexical relative frequencies[6] were created. To continue with the experiments we follow the directive, available on-line in open-source SAMT system,$(http://www.cs.cmu.edu//zollmann/samt)$ that consists of three parts:

1. Extraction of statistical translation rules from a training corpus: to extract purely lexical phrases by SOMAgent, which later were used to creat the grammar of the SAMT.
2. Cocke-Kasami-Younger (CKY+) [3] style chart-parser employing the statistical translation rules to translate test sentences.
3. A MER optimization and scoring tool (integrated into the chart parser) to tune the parameters of the underlying log-linear model on a held-out development corpus.

The target set of the training corpus was processing by the Penn Treebank parser of Charniak [4]. The size of the vocabulary of Penn Treebak is 61 elements.

We train the language model by using the beam-search decoder engine MER, in order to fit the weights of the characteristic functions and to generate the translations N-best and 1-best [21]. In the optimization process, the iterations number is limited to 10 and the 1000-best list was extracting. We used the measure BLUE like criterion of optimizations for maximize translation quality. Finally we did other sets of experiments with a phrase-based translation models using the same sentences but without preprocessing. The results for the system SAMT appear in the table 3.

Table 3 Evaluation of the translation for German to Spanish using SAMT

	BLUE
SAMT	42,20
SAMT-SOMAgent	63,11

5 Conclusions

The diagram described in the paper was created by using a MAS to apply a corpus preprocessing, which enabled the used of an open source SAMT.We applied the SomAgent to estimate the bilingual model. We experimented with different degree of linguistic analysis, from the lexical level to syntactic or semantic level, in order to generate a more precise alignment. Our work confirms the feasibility of the SOMAgent to automatically determinate the correct meaning of a word used in context and to collaborate in the use a word alignment to learn a phrase translation table. This approach confirms the idea that the linguistic information may be helpful, specially when the target language has a rich morphology (e.g. Spanish). Nevertheless this model offers a methodology that also illustrates the formation of a terminological mapping between two languages through an emergent conceptual space, and that can improve the first choice of the translator.

We have obtained interesting comparative results with regard to the measures BLUE: the SAMT system with SOMAgent overcomes his rival in 20 percent.

Acknowledgements. This work has been partially supported by the MICINN project TIN 2009-13839-C03-03.

References

1. Brown, P.F., Della Pietra, V.J., Della Pietra, S.A., Mercer, R.: The mathematics of statistical machine translation: parameter estimation. Comput. Linguist. 19(2), 263–311 (1993)
2. Casacuberta, F., Vidal, E., Vilar, J.M.: Architectures for speech-to-speech translation using finite-state models. In: Proceedings of the Workshop on Speech-to-Speech Translation: Algorithms and Systems, pp. 39–44 (2002)
3. Chappelier, C., Rajman, M.: A generalized CYK algorithm for parsing stochastic CFG. In: First Workshop on Tabulation in Parsing and Deduction (TAPD 1998), Paris, pp. 133–137 (1998)
4. Charniak, E.: A maximum entropyinspired parser. In: Proceedings of NAACL 2000, pp. 132–139 (2000)
5. Charniak, J.: Learning non-isomorphic tree mappings for machine translation. In: Proceedings of ACL 2003, (Compain Volume) pp. 205–208 (2003)
6. Chiang, D.: A hierarchical phrasebased model for statistical machine translation. In: Proceedings of ACL 2005, pp. 263–270 (2005)
7. Chiang, D.: Hierarchical phrase based translation. Computational Linguistics (2007)
8. Doddington, G.: Automatic evaluation of machine translation quality using n-gram cooccurrence statistics. In: Proceedings ARPA Workshop on Human Language Technology (2002)

9. Honkela, T.: Philosophical Aspects of Neural, Probabilistic and Fuzzy Modeling of Language Use and Translation. In: International Joint Conference on Neural Networks, IJCNN 2007, pp. 2881–2886 (2007)
10. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: NAACL 2003: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pp. 48–54. Association for Computational Linguistics, Morristown (2003)
11. Kohonen, T.: Self-organized Formation of Topologically Correct Feature Maps. In: Neurocomputing, pp. 511–522. The MIT Press, Cambridge (1990)
12. Kohonen, T.: Self-organized Maps. Proceedings of the IEEE 78(9), 1464–1480 (1990)
13. López, V., Alonso, L., Moreno, M.: A SOMAgent for Identification of Semantic Classes and Word Disambiguation. In: 7th International Conference on Practical Applications of Agents and Multi-Agent Systems (PAAMS 2009). Advances in Intelligent and Soft Computing, vol. 55, pp. 207–215 (2009) ISBN: 978-3-642-00486-5
14. Marcu, D., Wong, W.: A Phrase-Based, Joint Probability Model for Statistical Machine Translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, pp. 133–139 (2002)
15. Mariño, J.B., Banchs, R.E., Crego, J.M., Gispert, A., de Lambert, F.P., Costa-jussá, M.R.: N-gram based machine translation. Computational Linguistics 32(4), 527–549 (2006)
16. Melamed, I.D.: Statistical machine translation by parsing. In: Proceedings of ACL 2004, pp. 111–114 (2004)
17. Och, F., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics 29(1), 19–52 (2003)
18. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLUE: a method for automatic evaluation of machine translation. In: Proceedings of the Annual Meeting of the Association for Compuational Linguistics, ACL (2002)
19. Picó, D.: Combining Statistical and Finite-State Methods for Machine Translation. Thesis for the degree of doctor.Universitat Politécnica de Valéncia. Departament de Sistemes Informátics I Computació. Spain (2005)
20. Strube, V.L., Carneiro, P.R., Filho, I.: Distributing linguistic knowledge in a multiagent natural language processing system: re-modelling the dictionary. Procesamiento del lenguaje natura 23, 104–109 (1998)
21. Venugopal, A., Zollmann, A., y Vogel, S.: An Efficient Two-Pass Approach to Synchronous-CFG Driven Statistical MT. In: Proceedings of HLT/NAACL 2007, pp. 500–507 (2007)
22. Zollmann, A., Venugopal, A.: Syntax augmented machine translation via chart parsing. In: Proceedings of NAACL 2006 (2006)
23. Yamada, K., Knight, K.: A decoder for syntax-based statistical MT.In: Annual Meeting of the ACL. Proceedings of the 40th Annual Meeting on Association for Computational (2001)