Applied Soft Computing xxx (2010) xxx–xxx

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Short communication

A SomAgent statistical machine translation

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

Dept. Informática y Automática, University of Salamanca, Plaza de la Merced S/N, 37008, Salamanca, Spain

ARTICLE INFO

Article history:
Received 16 June 2009
Received in revised form 15 July 2010
Accepted 16 August 2010
Available online xxx

Keywords:
Natural language processing
Semantic Kohonen Maps
Automatic translator

ABSTRACT

The paper describes the process by which the word alignment task performed within SOMAgent works in collaboration with the statistical machine translation system in order to learn a phrase translation table. We studied improvements in the quality of translation using syntax augmented machine translation. We also 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 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 [19]. But there still persist several morphology and syntax errors, which derive from the inability of the model to handle word derivation, multi-word expressions, long-dependency syntax relationships, or semantic disambiguation, among other linguistic phenomena. Without providing this information to our models, it is only possible to rely on the indefinite increase in training data to improve current translation quality. Therefore, additional linguistic knowledge seems to be almost necessary.

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 [19,41] there appears the need to introduce alternative techniques to include information on morphology derivation and verb group information into word alignment algorithms. According to [41], the classification of verb forms can be improved for large data tasks by adding some granularity to the classes. For example, classifying all verb forms from a certain verb into more than just one class.

Most Natural Language Processing (NLP) systems traditionally use a sequential architecture that represents the classical linguistic levels. Previous studies have pointed to a distributed architecture as a means of dealing with this complex related information and making it available for text analysis. Some of them, such as [1,50,52] report research on the possibilities of using a multi-agent system (MAS) [56] in NLP, to represent cooperation among distinct linguistic levels.

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 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.

1.1. Self-Organizing Maps in natural language processing

The Self-Organizing Maps (SOMs) devised by Kohonen [23,24] are used for the extraction of information from a primary multidimensional signal and represent it in two dimensions. Much of the formal and computational study of written language is centered

* Corresponding author.
E-mail address: vivian@usal.es (V.F. López).
URL: http://dptoia.usal.es (V.F. López).

1568-4946/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.asoc.2010.08.018

on structural aspects, and not on semantics and pragmatics [17]. A SOM is used to resolve ambiguity in [49]. A model for lexical disambiguation is presented in [34]. The basic method for creating word category maps was introduced by Ritter and Kohonen [45,46]. Charniak [6] presented a scheme for grouping or clustering words into classes that reflect the commonality of some property. Pulkki [43] presents means for modeling ambiguity using SOMs. Contextual information has widely been used in statistical analysis of natural language corpora [17]. In recent years, this statistical approach has had considerable success, based on the availability of large parallel corpora and other methodological developments (consider e.g., [21,54]). Tikkala et al. [53] have presented connectionist models for simulating both normal and disordered word production as well as child language acquisition. The study in [28] indicates that connectionist modeling of language acquisition has made significant progress since Rumelhart and McClelland's pioneering model of the acquisition of the English past tense [47]. However, three major limitations need to be considered for the further development of neural network models of language acquisition:

First, some language acquisition models use artificially generated input representations that are isolated from realistic language uses.

Second, most previous models have used supervised learning through back-propagation as the basis for network training (see the models reviewed in [13,44]).

Third, neural network models of lexical learning [27] have not yet devised a method for modeling the incremental nature of lexical growth.

To address these three problems, a SOM neural network model of lexicon development was created. This model, referred to as DevLex, is designed to combine the dynamic learning properties of connectionist networks with the scalability of representation models. Previous work by [31], have shown that self-organizing neural networks, are particularly suitable as models of the human lexicon. Various aspects of modeling translation and language use have been considered in [17]. There has recently been considerable interest in the models of language evolution (see, e.g., [11]).

The remainder of this paper is structured as follows: the statistical machine translation is described in Section 2. The SAMT is presented in Section 3. The SOMAgent and our multi-agent system is further described in Section 4, the experimental work is reported in Section 5 and the conclusions are briefly outlined in Section 6.

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 [3]. When IBM researchers presented the statistical approach to MT, 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 [21] and Moses MT [59]. Marcu and Wong [32] 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 in the SMT model based on tuple N-grams [33], or Ngram-based SMT. This approach is an evolution of a previous Finite-State Transducer implementation of X-grams [4], which adapted speech recognition tools for speech-oriented MT. The result is a competitive SMT model whose basic unit is the tuple, composed of one or more words of the source language and for one or more words of the target language.

In the last year, much effort has 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 [35], and those based on the hierarchy of phrases [8].

3. Syntax augmented machine translation

Defined in [57] as a specific parameterization of the probabilistic synchronous context-free grammar (PSCFG) approach to MT, the syntax augmented machine translation takes advantage of nonterminal symbols used in monolingual parsing, to generalize beyond purely lexical translation. [9] 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 [7] 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 type:

$$X = \langle \gamma, \alpha, \iota, \psi \rangle \quad (1)$$

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 nonnegative 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 [3], or phrase-to-phrase rules [21]. However, they can be viewed as natural extensions to these well established approaches. [9] pointed out a procedure for learning PSCFG rules from word-aligned parallel corpora, using the phrase-pairs as a lexical basis for the grammar.

3.1. Phrase and SAMT rule extraction

Ref. [57] describe a process to generate a PSCFG given parallel sentence pairs and the use nonterminal labels learned from target language parse trees. The inputs to the SAMT rule extraction procedure are tuples, $\langle f, e, Phrases(a, f, e), \pi \rangle$, where f is a source sentence, e is a target sentence, a is a word-to-word alignment associating words in f with words in e, $Phrases(a, e, f)$, are the set of phrase pairs (source and target phrases) consistent with alignment a [21,39], and π is a phrase structure parse tree of e. SAMT rule extraction associates each phrase pair from $Phrases(a, e, f)$ with a left-hand-side label, and then applies the rule extraction procedure from [9] to generate rules with labeled nonterminal symbols. Consistently, all linguistic rules are included in the mapping table of phrases.

3.2. Rule generation

For the phrase translations on the parallel training data, the techniques and implementation described in [21] are used. This phrase table provides the purely lexical entries in the final hierarchical rule set that will be used in the decoding process. It then uses Charniak's parser [7] to generate the most likely parse tree for each

Please cite this article in press as: V.F. López, et al., A SomAgent statistical machine translation, Appl. Soft Comput. J. (2010), doi:10.1016/j.asoc.2010.08.018

target sentence in the training corpus. Next, it determines all phrase pairs in the phrase table whose source and target side occur in each respective source and target sentence pair. This defines the scope of the initial rules in the generation of synchronous context-free grammar (SynCFG) [10].

3.3. PSCFG decoding

The sentence specific grammars and language models are used in a bottom-up chart parsing decoder to perform the search in the probability space of the terminals for the target language. This is similar to a probabilistic context-free grammar and decoding is therefore an application of chart parsing, instead of the common method of converting the context-free grammar into Chomsky Normal Form and applying a Cocke–Kasami–Younger (CKY+) [5] that allows efficient decoding for grammars with more than two nonterminal symbols. The decoder integrates n-gram language models during its search, using the Cube Pruning algorithm described in [10] to mitigate the computational impact of this feature.

3.3.1. Minimum Error Rate training

The translation quality is represented by a set of the functions for every rule. These functions are trained via Minimum Error Rate (MER) [38] to maximize translation quality according to a user specified automatic translation metric, such as BLEU (Papineni et al.[40]) or NIST [12]. The weights of the functions are computed on the basis of the maximization of the BLEU measure.

4. Neural networks as non-parametric classification statistical tools

The relationships between neural networks and statistical methods have been recently analyzed [42]. In general terms neural networks have shown greater ability to classify than statistical tools. Moreover they do not need to satisfy the parametric assumptions of those techniques.

4.1. The Semantic memory of the SOMAgent model

In [18], the main focus is on modeling communities of conceptually autonomous agents. An agent is conceptually autonomous if it learns its representation of the environment by itself, where a concept is taken to be simply a means of specifying a relationship between language and world. Partial autonomy refers to a setting in which the learning process of an agent is influenced in some way by other agents. This influence can then serve as a basis for communication between agents. Thus, although each agent has an individual representation of the environment, the representations are related through the coordinating effort of communication between agents in situations where all agents have access to similar perceptions of the environment.

In our model, the environment consists of the context where the symbols (words) are represented during the process of learning, which implies that coded units should include a group of concurrent elements. In linguistics, the concept of the representation of the context is associated with a number of adjacent words. Thus similarity between words is a reflection of similarities of the context. The basic idea is to teach small context maps so that the SOMAgent [30] can process the contextual information into clusters. Each model vector of the single-word maps corresponds to a particular meaning of the word.

Our agent implements a mechanism of class analysis, i.e., clustering, to represent and identify groups of meanings that are semantically associated. A class is a set of associated meanings with a central concept, whose members can be concepts or other classes. The agent has been used to choose the correct meaning from various candidates. Therefore the agent is conceptually autonomous, but it has partial autonomy. For example, in cases where syntactic-semantic analysis is insufficient to solve a lexical ambiguity, the agent must collaborate with other agents and take the context into account.

Given the assumption that some sample data sets are mapped onto an array that will be called the map, the set of input samples is described by an n-dimension real vector $x(t)$. Each unit in the map contains an n-dimension vector $m(t)$. Let *Xs* be the vector which represents the symbolic expression of an element and *Xc* the representation of the context. The simplest neuronal model assumes that *Xs* and *Xc* are connected through the same neuronal unit, so that the vector X (the pattern) is formed by the concatenation of *Xs* and *Xc*:

$$X = \begin{bmatrix} X_s \\ X_c \end{bmatrix} = \begin{bmatrix} X_s \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ X_c \end{bmatrix} \quad (2)$$

The central foundation of the symbolic map is that the two parts have their own weights during the self-organization process. However, the size of the context predominates, reflecting the metric relationships of the members of the set, and implementing a spatial order that reflects semantic similarities.

To find semantic relationships between words, the semantic or conceptual space is explicitly modeled with the SOM algorithm [23]. The algorithm organizes the responses spatially in the map. The basic steps are:

1. Selection of the winning neuron.
2. Adaptation of the cell with the largest response (the winning cell), and of its topological neighbours, resulting from the current input.
3. Spatial concentration of the activity of the net in the cell (and optionally in its neighbours) that offers the largest response to the input.

Selection of the winning neuron. The learning algorithm iterates the following sequence:

1. Presentation of a given inputs to all cells.
2. Selection of the cell with the largest response to this input.

Let the vector which represents the actual input be $x = [x_1 \ldots x_n] \in R^n$ and the vector of weights for each cell, i, be $m_i = [m_{i1} \ldots m_{in}] \in R^n$. The criterion used to detect the cell that responds most is based on the Euclidean distance between x and m_i: choose the cell which is nearest and call it m_c, then $||x - m_c|| = min_i ||x - m_i||$.

Adaptation procedure. The weight vectors tend to approximate a form determined by the probability density function of the input vectors. Lateral interaction can be introduced by the definition of a group, N_c, of neighbouring cells around cell c. At each learning step, all the cells of N_c are updated while the rest remain unchanged. The adaptation process for the best m_i is defined by:

$$m_i(t+1) = \begin{cases} m_i(t) + \alpha(t)[x(t) - m_i(t)] & i \in N_c(t) \\ m_i(t) & i \notin N_c(t) \end{cases}$$

where $\alpha(t)$ is the *learning factor* $0 < \alpha(t) < 1$.

The overall architecture of multi-agent system is presented in Fig. 1. The SOMAgent receives perceptual inputs: linguistic expressions. There are potential actions: the agent can disambiguate an expression. The perception 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

Please cite this article in press as: V.F. López, et al., A SomAgent statistical machine translation, Appl. Soft Comput. J. (2010), doi:10.1016/j.asoc.2010.08.018

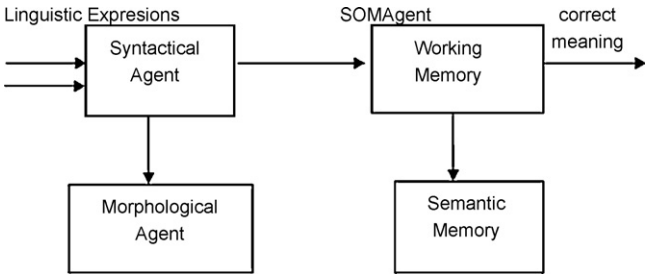

Fig. 1. Architecture of multi-agent system.

related to linguistic expressions: morphological, syntactical and semantic information about the lexical items that are necessary for the resolution of specific tasks.

4.2. The multi-agent system

The SOMAgent can be applied to find a map between vocabularies of two different languages in machine translation. According to [19], the mapping between any two languages is based on an intermediate level of representation. This knowledge is embedded in specific autonomous agents in a multi-agent system. In this section, we present the SOMAgent perspective from each of those agents. *Morphological agent*. Our lexical items are stored in a hierarchy of dictionaries:

1. Central dictionary or terms dictionary.
2. Idioms.
3. Terms that present lexical ambiguities.

Term description is carried out by the feature structure. The terms dictionary was constructed in declarative form, associating every term of the source language with its possible translations and corresponding morphological, syntactical and semantic information [29]. The conjugations of every term are generated through morphologic rules. Since German has a rich morphology inflection there are certain types of words that change their morphologic structure. The models allow termination groups to bind to bases that perform identical inflexion behavior, thus reducing the number of both characters and descriptors stored in the dictionary. The words appear in the dictionary in degree zero and not their declensions.

Syntactical agent. Generalized Phrase Structure Grammar (GPSG) [20] is used at this point. GPSG augments syntactic descriptions with semantic annotations that can be used to compute the compositional meaning of a sentence from its syntactic derivation tree. In order to implement this model, grammar knowledge comprising the initial tree models, which represent the structure of German sentences and the lexicalization dictionary form the syntactical agent knowledge. This agent can be seen as a subsociety [52], formed by agents handling simpler tasks 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: e.g., to assist in German Spanish translation. One possible organization for this subsociety is a group of autonomous agent handlers:

- Agent 1, initial trees.
- Agent 2, auxiliary trees.
- Agent 3, lexicalization dictionary.
- Agent 4, formalism operation and organization of the working memory of the subsociety.
- Agent 5, morphological and lexical transfer.

Table 1
English lexicon for the first application.

Words	Cat	Words	Cat
Mary–Peter–Jim	1	Horse–dog–cat	2
Beer–water	3	Lunch–foot	4
Runs–walks	5	Works–talks	6
Visits–telephones	7	Buys–sells	8
Likes–hates	9	Takes–eats	10
A lot–a little	11	Quickly–slowly	12
Frequently–rarely	13	Good–bad	14

Using German sentences as input, the parsing is performed, resulting in a decorated abstract syntax tree. The dictionary agent gets the morphological information from the morphological agent (agent 5). The dictionary agent must negotiate with agents 1–3. It sends the set of trees that must be evaluated to agent 4, who tests all possible combinations with the received information, and sets the values of the working memory.

The SOMAgent is implemented in C language under the UNIX operating system and using the Som-pack v3.1.10 software tool [25].

5. Data and experiments

Two examples are used to test the validity of the method proposed to demonstrate that the SOMAgent can be applied to the organization of linguistic information. The SOMAgent treats the organization of words into semantic classes according to their context, in a way that reflects a natural "organization".

5.1. Word classification with SOMAgent

5.1.1. Lexicon

The lexicon used in the implementation, shown in Table 1, is formed from words that are meaningful within a particular context (or domain), but it excludes words which are meaningless (i.e., they are independent of the domain or they belong to categories such as articles, prepositions, conjunctions and pronouns). This allows the net to be trained with a smaller range of errors.

These words define the type of context and comprise nouns, verbs and adverbs. Each class contains elements such as the name of a person, animals and inanimate objects. When taking [19] into account, the resulting view is called semantic holism. In a similar fashion, the SOM specifies a holistic conceptual space: "the meaning of a word is not based on some definition but is the emergent result of a number of encounters where a word is perceived or used in some context. Moreover, the emergent prototypes on the map are not isolated instances but they influence each other in the adaptive formation process".

5.1.2. Sentence patterns

To study semantic relationships in their pure form, it is recognised that semantic significance should not be inferred from any semantic pattern used for the encoding of individual words but only from the context where each word appears. Thus, in the simplest approach, all those words which occur in certain "windows" are represented by X_c and defined as inputs to the neural network. In this way vector inputs, X, to the network are created, by using the form of Eq. (2).

In the self-organizing process, the inputs consist of sequences of three words selected from certain patterns of classes shown in Table 1. Such class patterns are defined off-line (e.g. 1-5-12, 1-9-2, 2-5-14, ...). Sentences can then be constructed automatically by randomly selecting words from within each class. Two sample sentences for the class pattern 1-5-12 could be *"Peter runs quickly"* and *"Jim walks slowly"*.

Please cite this article in press as: V.F. López, et al., A SomAgent statistical machine translation, Appl. Soft Comput. J. (2010), doi:10.1016/j.asoc.2010.08.018

Table 2
Set of parameters used as input to the training software.

General – values	
Net dimension:	10×15
Topology:	hexagonal
Neighbourhood:	bobble
Initial weights:	at random
Organization phase	**Refinement phase**
Iterations: 5000	Iterations: 55,000
Radius: 3	Radius: 1
α: 00.1	α: 00.1

5.1.3. Training phase

The training phase consists of the sequential presentation of different semantically correct sentences and the adjustment of the weights described in Table 2 until the net converges. Each word in the vocabulary is encoded as a random 7-dimensional vector, according to a random number generator, and normalized to unit length. Starting from the existing vocabulary, groups of sentences are generated, each of which has a known meaning (as described in Table 1). Taking the class patterns, sentences are generated randomly. The sentences are concatenated into a stream and a word in this stream is selected at random. The context vectors are chosen and concatenated to make a 14 dimensional symbol vector, which is also normalized. The symbol vector is taken, multiplied by a length scale factor $a = 5$ (the relative influence of the symbol over the context). The input to the Kohonen net, with unsupervised learning, is used as the symbol vector concatenated with the corresponding context vector, according to the preceding formula (2).

The set of parameters used as input to the training software is shown in Table 2 and each "winning neuron" was labeled according to its corresponding word from the vocabulary.

5.1.4. Recognition phase

After training, the network becomes topologically ordered. We can verify that units of the map are actived for each input vector and then labeled. Fig. 2 shows how the resulting map separates the words according to their syntactic type e.g. verbs, nouns and adverbs which appear in separate zones; each of these zones is organized according to semantic similarities, forming "clusters" or semantic classes. These clusters are formed automatically. They should contain words that are semantically related, and should be maximal in the sense that if two terms are semantically related, then they will belong to the same class. For example adverbs which have opposite meanings appear next to each other on the map.

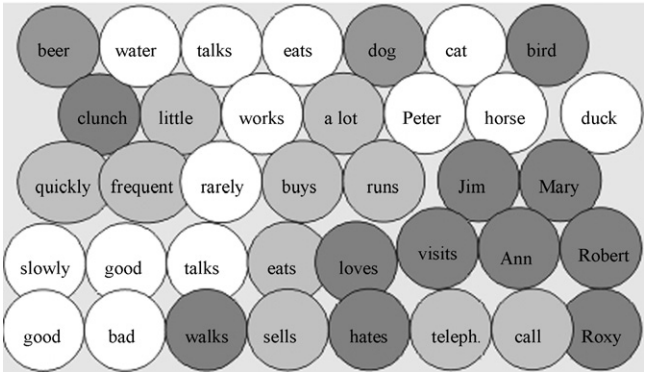

Fig. 2. The resulting map separates the words according to their syntactic type. Each of these zones is organized according to semantic similarities, forming "clusters".

For the analysis of the classes, lexical cohesion exists when words are used that repeat a conceptual category or that have semantic associations with one another [36]. These associations can be considered as a class of situations that cannot be predicted syntactically. Therefore, language parsers usually have to be able to separate meanings that share semantic associations. In this context, a class is defined as a set of written patterns that share a similar semantic context.

5.2. Word alignment with SOMAgent

In this section the SOMAgent 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 in building better translation models.

The alignment model as part of a whole translation scheme can also be defined as an independent NLP task. In fact, most new generation translation models treat word alignment as an independent result from the translation model. In [41] 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. In the last few years, much effort has been devoted to this matter [15,48,58] suggesting a combination of models based on shallow syntactic analysis (part-of-speech tagging and phrase chunking).

Our approach exploits the possibility of working with alignments at different levels of granularity, from the lexical to the semantic level, as [41], suggests. Therefore, assuming we are able to extract a set of tuples from a given parallel text, we can use the SOMAgent to estimate the bilingual model and to perform a corpus preprocessing for SMT in an Automatic German–Spanish Translator prototype. The aim of our linguistic agents is to participate in a society of entities with different skills, and to collaborate in word alignment to learn a phrase translation table. The most recently published methods on extracting a phrase translation table from a parallel corpus start with a word alignment.

For those cases where the society of linguistic agents is not sufficient to find the correct alignment and where contextual information is required to resolve ambiguity, the SOMAgent receives the linguistic expressions, and the semantic memory associates contextual information and gives the correct meaning.

5.2.1. Lexicon

In this case it is possible to create a parallel vocabulary set with all the words from the central dictionary that present lexical ambiguities and whose translations have to be determined from their context. A random vector is associates with each word. The vocabulary used for this specific example consists of nouns, ambiguous verbs and objects. The ambiguous verbs will define the fundamental context as indicated by the number on the right hand side. As shown in Table 3, the total number of contexts is 14.

The aim is to classify all the ambiguous verbs into more than just one class. The classification can be improved by adding some semantic granularity to the classes.

5.2.2. Sentences patterns

The Syntactical Agent divides the sentence into subject, verb, and object. The network inputs are enriched with features beyond the lexical ones, such as part-of-speech (PoS). The SVMTool [14,60] has been used for PoS-tagging and to provide data views at the word level (WP word and PoS).

The Kohonen net is trained with data linguistically annotated using the SOMAgent, with a large set of sentences that reflect every type of context in the corpus. Let us take, for example, a subset of German verbs that have *double meanings* and whose true mean-

Please cite this article in press as: V.F. López, et al., A SomAgent statistical machine translation, Appl. Soft Comput. J. (2010), doi:10.1016/j.asoc.2010.08.018

Table 3
German lexicon for the final application.

Words	Cat
Peter-Paul-Andreas	1
Klavier-Gitarre-Flöte	2
Fuβball-Karten-Schach	3
Film-Szene	4
Draht-Rohr-Stange	5
Feuer-Licht-Kerze	6
Durst	7
Programm-Kassette-Aufnahme	8
Schule-Kurs-Uni	9
Freund-Museum-Mutter	10
spielt	11
dreht	12
löscht	13
besucht	14

Table 4
Sentence patterns generated.

"Peter spielt Fuβball" (Peter plays football).
"Peter spielt Karten" (Peter plays cards).
"Peter spielt Schach" (Peter plays chess).

ing can only be selected from their context. The input for the map consist of theses words and their context.

5.2.3. Training phase

We consider the use of a subset of words in German, in a number of contexts from real-life situations. To illustrate the idea of using the SOM to find a mapping between ambiguous verbs from two different languages, we use the German verb "spielen" (to play) which has two meanings represented by different Spanish verbs: either "tocar", which appears in the context of playing a musical instruments *Klavier, Gitarre, Flöte (Cat = 2)*, or "jugar" which appears in the context of playing a game, *Fußball, Karten or Schach (Cat = 3)*.

Because the sentence patterns are generated based on the patterns of contexts shown in Table 3, sensible sentences are created covering every context. For example, with the pattern 1-11-3, a noun from context (1) the verb from context (11) and a game from context (3) are used. Sentences are created such as the ones shown in Table 4.

These sentences, by following the steps of the SOM algorithm [23], form a file of input data vectors for doing the training, creating the semantic memory with the semantic classes specified in Table 5.

5.2.4. Recognition phase

After training, the network becomes topologically ordered, and it is posible to verify what units of the map are activated for each input vector. The units are then labeled, with the principal semantic classes, taking the best answer for conducting via automatic model clustering to reduce the ambiguity.

In this way a trained net is created with the principal classes or with the active regions defined. A class is active if it contains any

Table 5
Semantic class for the network.

Class	Verb in Spanish	Verb in German
1	tocar	Spielt
2	jugar	Spielt
3	filmar	dreht
4	torcer	dreht
5	apagar	löscht
6	quitar	löscht
7	borrar	löscht
8	ir	besucht
9	visita	besucht

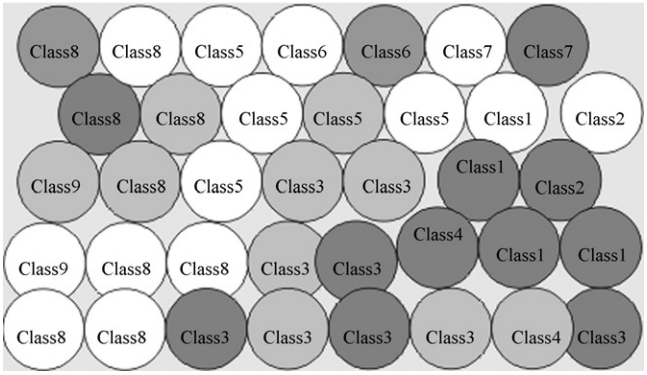

Fig. 3. Nine classes were created without any type of supervision. Each class represents a specific meaning in Spanish that corresponds to the same verb in German, which is close on the map.

of the meanings included in the training for ambiguous cases. Each class represents the meaning of the verbs according to the context. As shown in Fig. 3, nine classes were created, without any type of supervision. The resulting map separates the words according to their semantic similarities. Each class represents a specific meaning in Spanish that corresponds to the same verb in German, which is close on the map.

For those cases in which the SOMAgent is called on to collaborate in solving the ambiguity, it uses the results of the previous agents as input: the semantic agent searches for meanings associated with each word, forming key sentences with the combination of words in German which could not be disambiguated. These words are then fed into the network as input, thus allowing the network to classify each word within the active classes, taking the best answer as the correct meaning and the best alignment. We build a single translation model from the union of alignments from the data views and the SOMAgent work.

To illustrate, let us assume the case of word alignment possibilities illustrated in Table 6. For the sentence *Peter spielt Fußball* we take, the German verb *spielen* (to play) which 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 shown as both a verb and a noun. Considering these two words, which have the same lexical realization, the use of a single token adds noise to the word alignment process. The Syntactical Agent represents this information, by syntactic label (by means of linguistic data views) [16], as

Table 6
A case of word alignment possibilities on top of lexical units (A) and linguistic data (B).

					(A)
Peter	spielt	Fuβball,	Peter	spielt	Gitarre,
↙	↙	↙	↙	↙	↙
Peter	juega	futball,	Peter	toca	guitarra,
and	Peter	liebt	romantische	Spiele	
↙	↙	↙	↙	↙	
y	Peter	loves	romantic	plays	
					(B)
Peter	spielt	Fuβball,	Peter	spielt	Gitarre, (B)
NN	VBZ	NN	NN	VBZ	NN
↙	↙	↙	↙	↙	↙
Peter	juega	futball,	Peter	toca	guitarra,
PN	VB	NN	PN	VB	NN
and	Peter	liebt	romantische	Spiele	
CC	NN	VBZ	JJ	NNS	
↙	↙	↙	↙	↙	
y	Peter	(encantan)	románticas	obras	
CC	NN	VB	JJ	NNS	

Please cite this article in press as: V.F. López, et al., A SomAgent statistical machine translation, Appl. Soft Comput. J. (2010), doi:10.1016/j.asoc.2010.08.018

Table 7
Result of the aligned one.

German reference	Spanish reference
Peter spielt Fuβball	Peter juega futball.
Peter spielt Gitarre	Peter toca guitarra.

Table 8
Training set.

	Spanish	German
Sentences	40K	40K
Words	1.31	1.47
Length average	18.10	31.11
Vocabulary	41.12	21.10

spielen VBZ and *spielen NNS*. This allows us to distinguish between the two cases.

In Table 7 we can see the best alignment for the sentence *Peter spielt Fuβball*. The net finds the true meaning of the German verb *spielt*, aligning this entry inside the active classes, which in this case is class 2 (to play) whose meaning is *jugar* in Spanish. For the sentence *Peter spielt Gitarre*. The better alignment for *spielt* is *tocar* in Spanish (class 1).

A better translation for the sentence *Peter liebt romantische Spiele* is *Peter loves romantic plays* the better alignment for *romantische Spiele* is *obras románticas* in Spanish. Representing this verb, by syntactic label, as *spielen VBZ* and *spielen NNS* would allow us to distinguish between the two cases. By inspecting translation models we confirmed the better adjustment of probabilities.

The results of the previous examples demonstrate that the SOMAgent can be used to better estimate the bilingual model. Although this significant improvement in MT quality can be reported, taking into account the short time spent in the development of the linguistic tools using the SOM, where the tasks to determine the correct meaning of a word used in context emerge from the statistical properties of the training examples.

In addition, these translation models are smaller (between 50% and 60%) than the models based on lexical items alone. The reason is that we are working with semantic classes and the union of alignments from different data views, thus adding more constraints to the phrase extraction step.

5.3. Experimental work

We present the experimental results for a German to Spanish translation task, based on a set of sentences from the full DWDS corpus [62] of the news domain. The results were obtained using only the first 40K lines of the corpus. The statistical data set of the corpus can be seen in Table 8.

For phrase extraction we used Moses MT. There were 4.8M Moses style phrases that were extracted with the system. The first preliminary step requires the preprocessing of the parallel data using SOMAgent, so that the sentence is aligned and tokenized. The primary purpose is to deal with specific phenomena such as ambiguity and to generate more precise alignments. The tokenized output is formed from words that are meaningful within a particular context (or domain), but it excludes words that are meaningless because they are independent of the domain and belong to categories such as articles, prepositions, conjunctions and pronouns

The training data were provided for the sentences aligned (one sentence per line), in two files, one for the German sentences and one for the Spanish sentences. A phrase-based translation model was built from the output of the multi-agent systems to extract the purely lexical phrases, which were later used to create the grammar for the SAMT. Then, the script that forms part of the Moses MT System *grow-diag-final aligned* was run, and the word-to-word lexical relative frequencies [9] were created. To continue with the experiments we followed the directive that is available on-line in open-source SAMT system, [61], which consists of three parts:

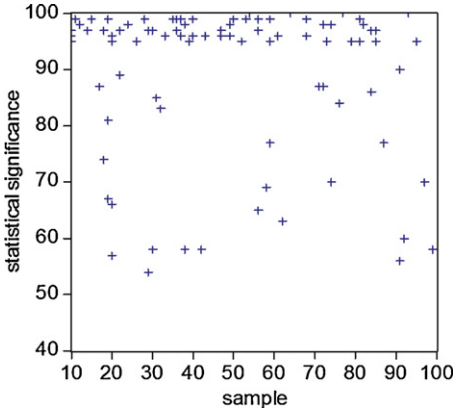

Fig. 4. Paired bootstrap resampling result on 100 samples. For 285 samples we draw the conclusion that SOMAgent preprocessing system is best with at least 97% statistical significance.

1. Extraction of statistical translation rules from a training corpus: to extract purely lexical phrases by SOMAgent, which were later used to create the grammar of the SAMT.
2. CKY+ style chart-parser employing the statistical translation rules to translate test sentences.
3. A minimum-error-rate 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 processed by the Charniak Penn Treebank parser [7]. The Penn Treebank has a vocabulary of 61 elements.

We trained the language model by using the MER beam-search decoder engine, which fit the weights of the characteristic functions and generates the translations *N*-best and 1-best [55]. In the optimization process, the iterations number was limited to 10 and the 1000-best list was extracted. Finally we performed other sets of experiments with a phrase-based translation model using the same sentences but without preprocessing. We used the BLEU measure as the criterion for optimizing the maximize translation quality.

5.3.1. Statistical significance

According to [22] the evaluation of machine translation systems has changed dramatically in the last few years. Instead of reporting human judgment of translation quality, researchers now rely on automatic measures, most notably the BLEU score. Since it has been shown that the BLEU score correlates with human judgment, an improvement in BLEU is taken as evidence of improvement in translation quality. But the BLEU measure does not lend itself to an analytical technique for assessing statistical significance; we use *the bootstrap resampling methods* [22] for this.

Using *the paired bootstrap resampling method* we can compare two systems. We translate the same test set with and without SOMAgent preprocessing, and measure the translation quality using the SOMAgent and not using it. The value of the scoring metrics lies in comparing the quality of the two different translation systems.

As in [22] we used a small collection of translated sentences, and repeatedly (1000 times) created new virtual test sets. We then

Please cite this article in press as: V.F. López, et al., A SomAgent statistical machine translation, Appl. Soft Comput. J. (2010), doi:10.1016/j.asoc.2010.08.018

Table 9
Evaluation of the translation from German to Spanish using SAMT-SOMAgent.

System	Bleu	Nist	mPer	mWer	Meteor
Baseline	43.20	9.16	36.89	49.45	58.50
SAMT	46.39	9.18	32.98	48.47	62.36
SAMT-SOMAgent	48.00	9.35	33.20	47.54	62.27

perform experiments using the BLEU score to compare both systems. Results are displayed in Fig. 4. For each set, we compute the evaluation metric score for both systems and note the best system.

We estimate statistical significance for 100 different test sets with 300 sentences each (the same test samples used in previous experiments). For 285 samples we draw the conclusion that the SOMAgent preprocessing system is the best with at least 97% statistical significance. The collection of translated sentences with SOMAgent preprocessing system is statistically different from the collection of translated sentences without SOMAgent, especially BLEU (48% vs. 46%).

The BLEU score difference on the 300 sentence test set is 2% (see to Table 9). According to [22] a small 300 sentence test set is often sufficient to detect the superiority of one of the systems with statistical significance. Even for small test sets of 300 sentences, we can reliably draw the right conclusion, if the true BLEU score difference is at least 2–3%.

Finally, we compare the result for the same set of tests carried out using the same tools (automatic measures) with and without SOMAgent preprocessing. Table 9 presents MT results for the test set for the German-to-Spanish task for both variants. It is compared to a baseline variant based only on lexical items [16].

For our final evaluation we selected a set of two classic metrics, BLEU and NIST, and the variants corresponding to different families: mPER [26], mWER [37] and METEOR [2].

In the case of SAMT-SOMAgent all metrics significantly outperform the baseline and SAMT system. We suspect this may be because the SOMAgent generates more precise alignments from different data views with linguistic knowledge according to their context (semantic classes).

6. Conclusions

The diagram described in the paper was created using a MAS to apply a corpus preprocessing, which enabled the use of a quality open source SAMT. We applied the SomAgent to estimate the bilingual model and experimented with different degrees 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 determine the correct meaning of a word used in context and to collaborate in the use of a word alignment to learn a phrase translation table.

This approach confirms the idea that the linguistic information may be helpful, especially when the target language has a rich morphology (e.g. Spanish). Nevertheless, with regard to the computational cost, the SAMT system with SOMAgent gives poorer results. However, 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.

References

[1] J. Balsa, G. Lopes, A distributed approach for a robust and evolving NLP system, Lecture Notes in Computer Science, Springer, Proceedings of the NLP 2000 Conference, ISBN 978-3-540-67605-8, 2000, pp. 151–161.
[2] S. Banerjee, A. Lavie, METEOR: an automatic metric for MT evaluation with improved correlation with human judgments, in: Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005.
[3] P.F. Brown, V.J. Della Pietra, S.A. Della Pietra, R. Mercer, The mathematics of statistical machine translation: parameter estimation, Comput. Linguist. 19 (2) (1993) 263–311.
[4] F. Casacuberta, E. Vidal, J.M. Vilar, Architectures for speech-to-speech translation using finite-state models, in: Proceedings of the Workshop on Speech-to-Speech Translation: Algorithms and Systems, 2002, pp. 39–44.
[5] C. Chappelier, M. Rajman, A generalized CYK algorithm for parsing stochastic CFG, in: In First Workshop on Tabulation in Parsing and Deduction (TAPD98), Paris, 1998, pp. 133–137.
[6] E. Charniak, Statistical Language Learning, MIT Press, Cambridge, MA, 1993.
[7] E. Charniak, A maximum entropyinspired parser, in: En Proceedings of NAACL 2000, 2000, pp. 132–139.
[8] E. Charniak, Learning non-isomorphic tree mappings for machine translation, in: En Proceedings of ACL 2003 (companion volume), 2003, pp. 205–208.
[9] D. Chiang, A hierarchical phrase based model for statistical machine translation, in: En Proceedings of ACL 2005, 2005, pp. 263–270.
[10] D. Chiang, Hierarchical phrase based translation, Comput. Linguist. (2007).
[11] M. Christiansen, S. Kirby, Language Evolution, Oxford University Press, 2003.
[12] G. Doddington, Automatic evaluation of machine translation quality using n-gram cooccurrence statistics, in: Proceedings ARPA Workshop on Human Language Technology, 2002.
[13] J. Elman, A. Bates, A. Johnson, A. KarmiloffSmith, D. Parisi, K. Plunkett, Rethinking Innateness: A Connectionist Perspective on Development, MIT Press, Cambridge, MA, 1996.
[14] J. Giménez, L. Márquez, SVMTool: a general POS tagger generator based on Support Vector Machines, in: Proceedings of 4th LREC, 2004.
[15] J. Giménez, L. Márquez, Combining linguistic data views for phrase-based SMT, in: Proceedings of the Workshop on Building and Using Parallel Texts, ACL, 2005.
[16] J. Giménez, L. Márquez, The LDV-COMBO system for SMT, in: Proceedings of the Workshop on Statistical Machine Translation, New York City, Association for Computational Linguistics, 2006, pp. 166–169.
[17] T. Honkela, Self-Organizing maps in natural language processing, Thesis for the degree of Doctor of Philosophy, Helsinki University of Technology, Department of Computer Science and Engineering, Public defense at 12th of December, 1997.
[18] T. Honkela, J. Winter, Simulating language learning in community of agents using self-organizing maps. Technical report, Helsinky University of Technology, Computer and Information Science Report A71. ISBN 951-22-6881-7, 2003.
[19] T. Honkela, Philosophical aspects of neural, probabilistic and fuzzy modeling of language use and translation, in: International Joint Conference on Neural Networks, IJCNN, 2007, 2007, pp. 2881–2886.
[20] G. Gazdar, E. Klein, G.K. Pullum, I.A. Sag, Generalized Phrase Structure Grammar. Cambridge, MA: Harvard University Press and Oxford: Basil Blackwell's, 1985.
[21] P. Koehn, F.J. Och, D. Marcu, Statistical phrase-based translation, in: NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Morristown, NJ, USA: Association for Computational Linguistics, 2003, pp. 48–54.
[22] P. Koehn, Statistical significance tests for machine translation evaluation, in: En Proceedings of EMNLP, 2004, pp. 388–395.
[23] T. Kohonen, Self-organized Formation of topologically correct feature maps. Neurocomputing, The MIT Press, Cambridge, 1990, pp. 511–522.
[24] T. Kohonen, Self-organized maps, Proc. IEEE 78 (9) (1990) 1464–1480.
[25] T. Kohonen, SOM-PAK: The self-Organizing Map Program Package, Helsinki University of Technology Laboratory of Computer and Information Science, Finland, 1995.
[26] G. Leusch, N. Ueffing, H. Ney, A novel string-to-string distance measure with applications to machine translation evaluation, in: Proceedings of MT Summit IX, 2003.
[27] P. Li, Language acquisition in a self-organizing neural network model, in: P. Quinlan (Ed.), Connectionist Models of Development: Developmental Processes in Real and Artificial Neural Networks, Psychology Press, New York, 2003, pp. 115–149.
[28] P. Li, Farkas, Early lexical development in a self-organizing neural network, Neural Networks 17 (2004) 1345–1362.
[29] V.F. López, Desambiguación semántica basada en métodos conexionistas para un problema de traducción automtica Alemán Español, Thesis for the degree of Doctor in computer science, Valladolid University, Spain, 1996.
[30] V.F. López, L. Alonso, M. Moreno, A SOMAgent for identification of semantic classes and word disambiguation. 7th International Conference on Practical Applications of Agents and Multi-Agent Systems (PAAMS'09), Series: Advances in Intelligent and Soft Computing, vol. 55/2009. ISBN: 978-3-642-00486-5, 2009, pp. 207–215.
[31] B. MacWhinney, Lexicalist connectionism, in: P. Broeder, J.M. Murre (Eds.), Models of Language Acquisition: Inductive and Deductive Approaches, Oxford University Press, Oxford, UK, 2001, pp. 932–942.
[32] D. Marcu, W. Wong, A phrase-based, joint probability model for statistical machine translation, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, 2002, pp. 133–139.
[33] J.B. Mariño, R.E. Banchs, J.M. Crego, A. Gispert, P. Fonollosa, M.R. Costa-jussá, N-gram based machine translation, Comput. Linguist. 32 (December 4) (2006) 527–549.

Please cite this article in press as: V.F. López, et al., A SomAgent statistical machine translation, Appl. Soft Comput. J. (2010), doi:10.1016/j.asoc.2010.08.018

[34] M.R. Mayberry III, R. Miikkulainen, Lexical disambiguation based on distributed representations of context frequency, in: Proceedings of the 16th Annual Conference of the Cognitive Science Society, 1994.
[35] I.D. Melamed, Statistical machine translation by parsing, in: Proceedings of ACL 2004, 2004, pp. 111–114.
[36] G. Morris, Hirst, Semantic interpretation and ambiguity, Artif. Intell. 34 (1988) 131–177.
[37] S. Nieben, F.J. Och, G. Leusch, H. Ney, Evaluation tool for machine translation: fast evaluation for MT research, in: Proceedings of the 2nd International Conference on Language Resources and Evaluation, 2000.
[38] F. Och, H. Ney, A systematic comparison of various statistical alignment models, Comput. Linguist. 29 (1) (2003) 19–52.
[39] F. Och, H. Ñey, The alignment template approach to statistical machine translation, En Comput. Linguist. (2004).
[40] K. Papineni, S. Roukos, T. Ward, W.J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 2002.
[41] D. Picó, Combining statistical and finite-state methods for machine translation, Thesis for the degree of doctor, Universitat Politècnica de València, Departament de Sistemes Informátics I Computació, Spain, 2005.
[42] A. Pitarque, J.C. Ruiz, J.F. Roy, Las redes neuronales como herramientas estadsticas no paramétricas de clasificacin, Psicothema ISSN 0214-9915 CODEN PSOTEG Vol. 12, Supl. no. 2, 2000, pp. 459–463.
[43] V. Pulkki, Data averaging inside categories with the selforganizing map. Report A27, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland, 1995.
[44] P. Quinlan, Modeling human development: In brief, in: P. Quinlan (Ed.), Connectionist Models of Development: Developmental Processes in Real and Artificial Neural Networks, Psychology Press, New York, 2003, pp. 112–121.
[45] H. Ritter, T. Kohonen, Self-organizing semantic maps, Biol. Cybern. 61 (4) (1989) 241–254.
[46] H. Ritter, T. Kohonen, Learning 'semantotopic maps' from context, in: Proceedings of IJCNN-90-WASH-DC, International Joint Conference on Neural Networks, vol. I, Lawrence Erlbaum, Hillsdale, NJ, 1990, pp. 23–26.
[47] D. Rumelhart, J. McClelland, Parallel Distributed Processing, vol. 2: Psychological and Biological Models, Chapter on Learning the Past Tenses of English Verbs, MIT Press, Cambridge, MA, 1986, pp. 216–271.
[48] C. Schafer, D. Yarowsky, A two-level syntax-based approach to Arabic-English statistical machine translation, in: Workshop on Machine Translation for Semitic Languages, New Orleans, Louisiana, 2003.
[49] J.C. Scholtes, Resolving linguistic ambiguities with a neural data-oriented parsing (DOP) system, in: I. Aleksander, J. Taylor (Eds.), Artificial Neural Networks, 2 vol. II, North-Holland, Amsterdam, Netherlands, 1992, pp. 1347–1350.
[50] J.L. Silva, V. Lima, An alternative approach to lexical categorical desambiguantion using a multi-agent system architecture, in: Proceedings of the RANLP'97, Bulgaria, 1997, pp. 6–12.
[52] V.L. Strube, P.R. Carneiro, I. Filho, Distributing linguistic knowledge in a multiagent natural language processing system: re-modelling the dictionary. Procesamiento del lenguaje natural, ISSN 1135-5948, No. 23, 1998, pp. 104–109.
[53] A. Tikkala, et al., The production of finnish nouns: a psycholinguistically motivated connectionist model, Connect. Sci. 9 (3) (1997) 295–314.
[54] R. Zhang, H. Yamamoto, M. Paul, H. Okuma, K. Yasuda, Y. Lepage, E. Denoual, D. Mochihashi, A. Finch, E. Sumita, The NiCTATR statistical machine translation system for IWSLT, in: Proceedings of the International Workshop on Spoken Language Translation, Kyoto, Japan, 2006, pp. 83–90.
[55] A. Venugopal, A. Zollmann, S. Vogel, An efficient two-pass approach to Synchronous-CFG Driven Statistical MT, in: En Proceedings of HLT/NAACL, 2007, pp. 500–507.
[56] M. Wooldridge, Intelligent agents, in: G. Weiss (Ed.), Multiagent Systems, The MIT Press, 1999.
[57] A. Zollmann, A. Venugopal, Syntax augmented machine translation via chart parsing, in: Proceedings of NAACL 2006, 2006.
[58] K. Yamada, K. Knight, A decoder for syntax-based statistical MT, in: Annual Meeting of the ACL, Proceedings of the 40th Annual Meeting on Association for Computational, 2001.
[59] http://www.statmt.org/moses/.
[60] http://www.lsi.upc.es/nlp/SVMTool/.
[61] http://www.cs.cmu.edu/zollmann/samt.
[62] http://utils.mucattu.com/.

Vivian F. López (PhD). Received a PhD in Computer Science from the University of Valladolid in 1996. At present she is Associate Professor at the University of Salamanca (Spain) where she has been since 1998. Member of the Data Mining Group (http://mida.usal.es/) at the University of Salamanca (Spain). She has done research on natural language processing and neural networks. She has also 70 papers published in recognized journals, workshops and symposiums, 16 books and book chapters and 20 technical reports most of these in this topics.

Juan M. Corchado (PhD). Received a PhD in Computer Science from the University of Salamanca in 1998 and a PhD. in Artificial Intelligence (AI) from the University of Paisley, Glasgow (UK) in 2000. At present he is Dean at the Faculty of Computer Sciences, Associate Professor, Director of the Intelligent Information System Group (http://bisite.usal.es) and Director of the MSc programs in Computer Science at the University of Salamanca (Spain), previously he was sub-director of the Computer Science School at the University of Vigo (Spain, 1999–00) and Researcher at the University of Paisley (UK, 1995–98). He has been a research collaborator with the Plymouth Marine Laboratory (UK) since 1993. He has leaded several Artificial Intelligence research projects sponsored by Spanish and European public and private institutions and has supervised seven PhD. students. He is the co-author of over 130 books, book chapters, journal papers, technical reports, etc. published by organizations such as Elsevier, IEEE, IEE, ACM, AAAI, Springer Verlag, Morgan Kaufmann, etc., most of these present practical and theoretical achievements of hybrid AI and distributed systems. He has been President of the organizing and scientific committee of several international symposiums.

Juan Francisco de Paz (PhD Student). At this moment, he is completing his studies of PhD in Computer Science at University of Salamanca (Spain). He obtained a Technical Engineering in Systems Computer Sciences degree in 2003, an Engineering in Computer Sciences degree in 2005 at the University of Salamanca and at this moment he is finishing Statistic in the same university. He has been co-author of published papers in several journals.

Sara Rodríguez (PhD Student). Currently, she is pursuing her studies of PhD in Computer Science at University of Salamanca (Spain). She obtained a Technical Engineering in Systems Computer Sciences degree in 2004, an Engineering in Computer Sciences degree in 2007 at the University of Salamanca. She has participated as a co-author in papers published in recognized international conferences and symposiums.

Javier Bajo (PhD). Received a PhD in Computer Science and Artificial Intelligence from the University of Salamanca in 2007. At present he is Associate Professor at the Pontifical University of Salamanca (Spain). He obtained an Information Technology degree at the University of Valladolid (Spain) in 2001 and an Engineering in Computer Sciences degree at the Pontifical University of Salamanca in 2003. He has been member of the organizing and scientific committee of several international symposiums such as CAEPIA, IDEAL, HAIS, etc. and co-author more than 100 papers published in recognized journals, workshops and symposiums.

Please cite this article in press as: V.F. López, et al., A SomAgent statistical machine translation, Appl. Soft Comput. J. (2010), doi:10.1016/j.asoc.2010.08.018