Computer Engineering and Intelligent Systems
ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)
Vol.4, No.10, 2013
www.iiste.org

Developing an Architecture for Translation Engine using Ontology

Faten Kharbat
Department of Management Information System
Al Ain University of Science and Technology, Abu Dhabi, UAE
E-mail of the corresponding author: faten.kharbat@aau.ac.ae

Abstract
In translation, analyzing the input sequence in order to determine its grammatical structure with respect to the given formal grammar is called the parsing procedure (Bataineh & Bataine, 2009). In this research, the main idea of the proposed architecture is to utilize the WordNet ontology to be the syntactic guide along with the Transition Network Grammar to determine the grammatical structure for the text to be translated. This is followed by a mapping process between the source and target languages which will enhance the accuracy of the result. Also, it will guarantee that the output will be syntactically acceptable according to the target language rules. This research is an open research which is having ongoing results and developments. Herein, the main architecture is described to open the door for several future steps for further integration with other techniques and approaches.
Keywords: Translation, WordNet, Transition Network Grammars, mapping engine, parsing procedure.

1. Introduction

Machine Translation is one of the open problems that need to be addressed in different way each time the science develops its techniques and has new approaches. Ontologies (Gruber, 1993) and Transition Network Grammars (Woods, 1970) are two techniques that can be integrated to develop a new architecture in order to have a better accuracy in machine translation. This research describes the overall proposed architecture to illustrate the basic techniques and their integration together without going into specific details. An example is described at the end of the paper to give a clearer picture.

The paper is divided into 5 sections. A background is described in the section 2 to give a brief overview of all the used elements within the research. Section 3 describes the overall proposed architecture, and section 4 gives a brief general example to translate a sentence from English to Arabic using the proposed architecture. Finally, section 5 concludes and outlines some of the future directions.

2. Background

2.1 Machine Translation

Machine Translation (MT) is known as the automatic translation from an existing text written in a one natural language (called source language) into an equivalent text in different language (called the target language) using computers (Lopez, 2008)(Goutte et al., 2009). The translation process is divided into two main steps (ML, 2010): decoding the source text and encoding the text in the target language. One of the good overviews about machine translation can be found in Hutchins (2003).

Many data mining techniques have been used to solve such problem. One of such techniques is the Rule-based Machine Translation (RBMT) based on linguistic information about source and target languages, such as dictionaries and grammars covering the main semantic, morphological, and syntactic regularities of each language. In a RBMT text is parsed, an intermediary is created, and the text in the target language is generated. An example of such method can be found in (Wong et al., 2006).

On the other hand, Statistical-based Machine Translation (SMT) try to generate translations using statistical methods based on bilingual text corpora such as (Rodríguez et al., 2008) and (Hwang et al., 2007). SMT can be categorized into three main categories: word-based machine translation, phrase-based machine translation, and syntax-based machine translation. A full explanation can be found in (Goutte et al., 2009).

Example-based Translation Machine (EBTM) can be considered as an implementation of case-based reasoning in machine learning by using bilingual corpus as its main knowledge base at run-time (Güvenir & Cicekli, 1998). EBMT is suitable in phrasal verbs which have highly context-dependent meanings. Phrasal verbs produce specialized context-specific meanings that may not be derived from the meaning of the constituents. There is almost always an ambiguity during word-to-word translation from source to the target language.

Finally, there are some hybrid approaches that mix between concepts and existing approaches. One example is the HMT system (Boretz, 2009) which integrated statistical and rule-based translation methodologies. HMT is believed that it "improves translation of large volumes of speech and text compiled from a variety of sources and assists linguists, translators, and analysts in achieving greater productivity more quickly and in a more cost effective manner" (Boretz, 2009).

Many applications/systems have been developed to enhance machine translation, some of them are online and

others are stand alone application under specific operating system. However, the most famous online translation service are BabelFish (http://www.babelfish.org/), Google translation tool (http://translate.google.com), and Windows Live Translation (http://www.microsofttranslator.com/). In this research Google tool is been used to test simple sentence's translation from English to Arabic.

2.2 Recursive Transition Network Grammar

Transition Network Grammars is a finite-state automata that represent transcriptions of the rules of a context-free grammar which is considered as a directed graph with labelled states and arcs (Woods, 1970). Recursive Transition Network (RTN) Grammars is a development for the finite state automata with recursive complexion which is used to parse the syntax of natural language phrases (Bataineh & Bataine, 2009).

Each language has its own properties and features that can be presented clearly by its own RTN. For example, (Woods, 1970) built a simple network to describe the structure of the English language. Also, (Bataineh & Bataine, 2009) used the RTN to develop an Arabic Parser with the aim of analyzing and extracting the attributes of Arabic words.

2.3 WordNet Ontology

In the last decade, ontologies have been considered as the backbone technology in most knowledge-based applications. As ontologies have become more common, their applicability has ranged from artificial intelligence areas such as knowledge representation and natural language processing to different fields such as information integration and retrieval systems, requirements analysis, and lately in semantic web applications.

The most popular definition of ontology was proposed by Gruber (1993), who defined it as "…a formal, explicit specification of a shared conceptualisation". In this definition, Gruber placed emphasis on formalising the specification of concepts and relations, which in turn allows for knowledge representation and sharing among different agents. Studer et al. (1998) analysed this definition, and identified four main concepts: formal, explicit, shared, and conceptualisation. The term formal means that an ontology should be machine readable; explicit implies that all concepts and constraints used are explicitly defined; shared indicates that an ontology should capture consensual knowledge accepted by the communities involved; and conceptualisation refers to an abstract model of phenomena in the real world arrived at by identifying the relevant concepts of those phenomena. Another relevant definition of an ontology was introduced by Guarino (1998): "a set of logical axioms designed to account for the intended meaning of a vocabulary". In this definition, Guarino highlighted the role of logic theory as a means of representing an ontology.

As a conclusion, ontologies formalize the semantics of the domain explicitly by describing their elements; and thus, they consist of concepts that describe the internal features of the concepts, and the properties that describe the relationships between these concepts. Ontologies are based on a shared and consensual domain knowledge agreed by a community.

Different kinds of ontologies exist that have been specified for different application domains thereby representing different types of knowledge. As a knowledge representation, WordNet (Miller, 1990) is a semantic network with a core concept called *synset* and 10-20 primitive relations (such as hyponym, hypernym, and antonym) and about 200,000 nodes (OntologySummit, 2007). Each node corresponds to a synset, which is a set of synonymous natural language words that stands for a single word sense. Each word sense has a definition in natural language like a dictionary. This can be imagined to be as a dictionary with thesaurus, organized around synsets / word senses. The synsets are sets of synonyms which gather lexical items having similar significances. A list of enumerates of the semantic relations available in WordNet can be found in details in (Elberrichi et al., 2008). WordNet is considered as an ontology in which each word sense nodes is a concept, entity types, or classes.

WordNet has been used in many applications and studies such as Text Categorization (Elberrichi et al., 2008), automatic generation of concept hierarchies (Lee et al., 2008), in health sector (Fellbaum et al., 2006) ...etc. Regarding translation, WordNet has been used in different ways such as in Salam et al. (2008) which used the WordNet in an example based English-Bengali MT.

3. The Proposed Architecture

Usually, the procedure of translating needs external information to ensure generating a reasonable structure of the blocks of the target text. Parsing is the process of analyzing an input sequence in order to determine its grammatical structure with respect to a given formal grammar (Bataineh & Bataine, 2009). The main idea of the proposed architecture is to utilize the WordNet ontology to be the syntactic guide along with the Transition Network Grammar to determine the grammatical structure for the text to be translated. After determining the grammatical structure in the source text, a mapping step is to be executed to ensure that the generated text will be syntactically correct with the feature of being readable and very close to human translation.

Within this research, the structure (grammar) rules are represented by a finite-state automaton of a Transition Network Grammar. They guide syntax analysis and generate a syntactical accepted parse tree. If a sentence is

syntactical correct, then it should be accepted by the Transition Network Grammars via parsing the network comprises of these arcs. Using the fact that a word may have different syntactical categories, the benefit of the WordNet ontology is to determine the syntax analysis of sentences to discover whether the sentence's structure is acceptable or not.

The proposed architecture is illustrated in Figure 1. It consists of two main stages: syntax analysis and mapping engine. The input for the syntax analysis stage is the source text (i.e., English in our example) which will be parsed via the Transition Network Grammar guided by the WordNet ontology. This stage will analyze the sequence of tokens to determine the grammatical structure with respect to the source language. Some more technical details can be found in Hamaydeh & Kharbat (2009). The output of this stage will be a full parsed text with clear and correct structure which can be formed as a parsed tree.

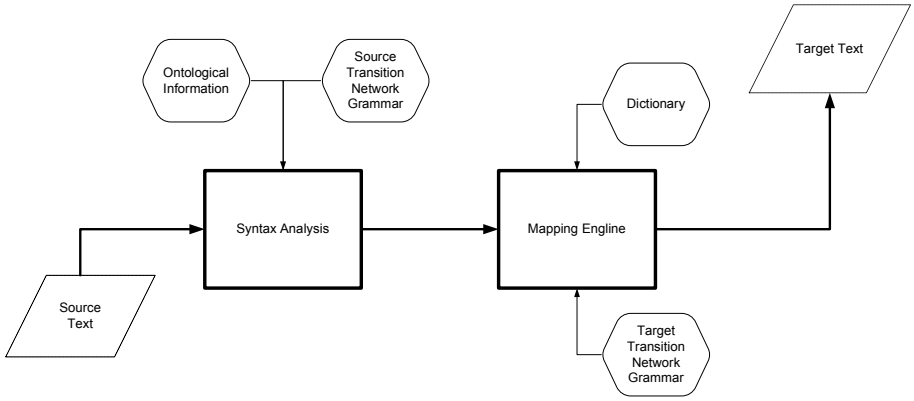
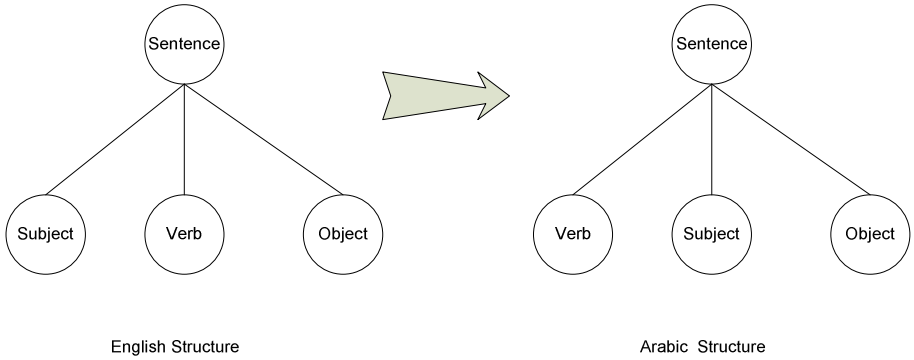


Figure 1. The proposed architecture.

This parsed tree is used as an input into the mapping engine which is the next phase. The second stage addresses the differences between the source and target languages in terms of the structure of the sentences and prefixes and postfixes used to structure the words. The mapping engine is designed to make use of the Transition Network Grammar for the target language (such as (Bataineh & Bataine, 2009) for the Arabic language) to reconstruct the sentences in the correct way after translating the words and phrases in the previous stage.

The main idea of the mapping engine is to find a compatible composition between the source and the target languages. The input parse tree will be used to generate an equivalent Transition Network Grammar for the target language. Figure 2 shows a simple example of a compatible parse trees for English and Arabic sentence. The subject in an English sentence comes before the verb, but in Arabic the verb should always precede the subject.

Figure 2. Simple example for a compatible parse tree between English and Arabic



In this stage the usual process in any translation process occurs, that is the actual translation of words and phrases into the target language (i.e., Arabic language in our example). This can be done by using a dictionary or a version of the WordNet in the target language.

4. An Example

In this section, a simple example is taken to illustrate the idea of the proposed architecture. The example is to translate an English sentence into an Arabic one and to compare the results with some other translation tools such as Google translation. The sentence that is chosen may have some oblique words with respect to semantic and grammar issues.

The sentence is: “He can can the can”.

It is clear that the first “CAN” in the sentence does not have the same lexical and syntactic analysis as the second “CAN”. Also, it can be simply seen that the third “CAN” is a noun rather than a verb.
If this statement is entered in Google translation service the result will be:
“ويمكن في وسعه العلبة” which indicates a real problem dealing with verbs and nouns! In another words, Google treated the “CAN”s with misleading quality of translation. However, if the sentence follows the proposed structure, the stages will be as follows:

Figure 3. Google translation of the sample sentence

The active sentence in English can be divided into three parts:

Nominal Part + Verbal Part + Complement Part

Each part has a separate Finite Automata. Figure 4 shows a part of the Nominal Part automata which illustrates that a nominal part may contain a noun only, or it may consists of noun, adjective and noun, ...etc.

Figure 4. A small segment from the nominal part automata

The first part of the sentence can be parsed to be a subject pronoun, which is represented by the arc between node 0 and node 1 in Figure 4.

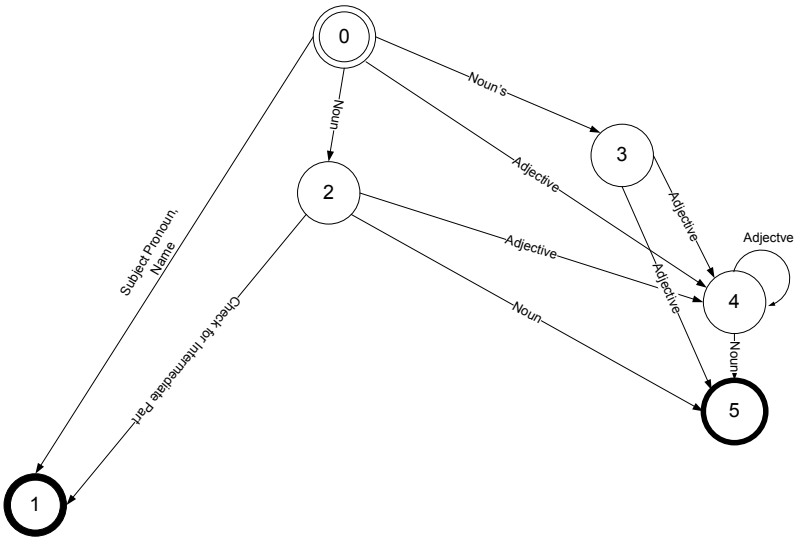


Figure 5 shows a small part from the Verbal Part finite automata, which illustrate that a verbal part may consists of only a verb, or an auxiliary verb followed by a verb, ...etc.

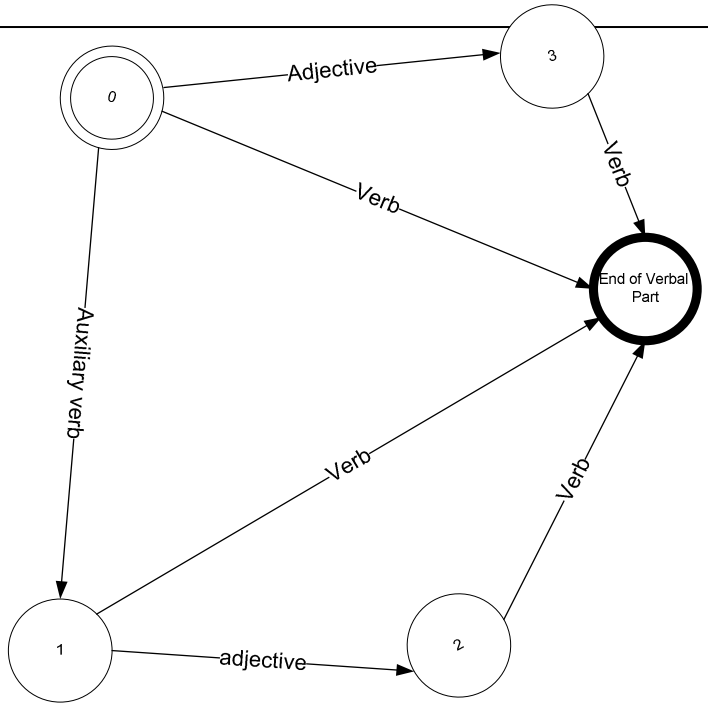


Figure 5. A small segment from the verbal part automata

Regarding the example taken in this section, the sentence is parsed via WordNet and the word “can” is found to have a verb sense and a noun one as follows:

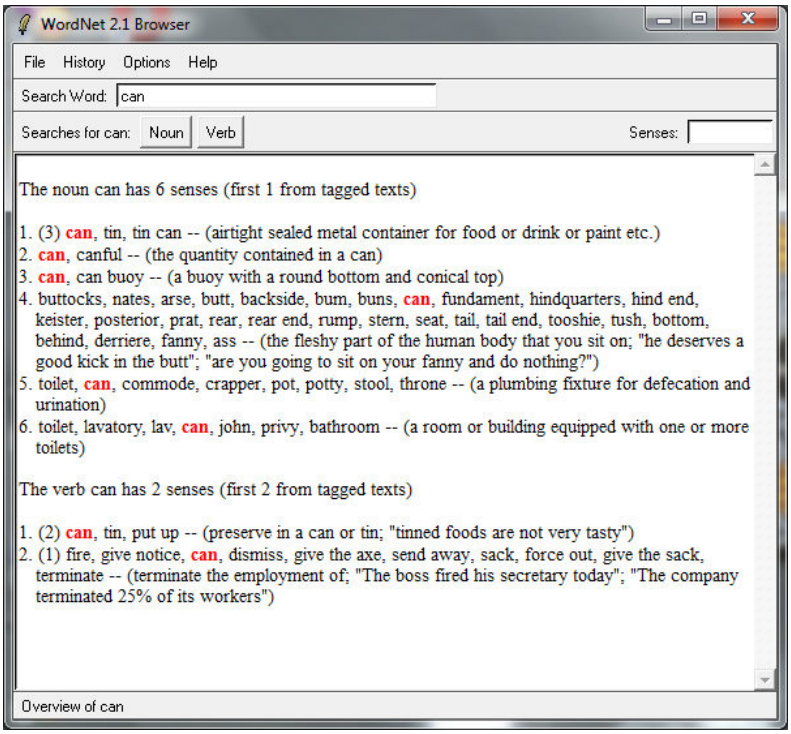


Figure 6. The analysis of the word "CAN" in the WordNet

As per Figure 5, the word "CAN" may be a verb, but the second "CAN" indicates that the path should be taken is: node 0, node 1, and end of verbal part. This will help in determining the most suitable structure in the target language; therefore, more accurate translation will be provided.

The same procedure will be applied to the complement part "the can", which is illustrated in Figure 7 to be a noun (the arc between node 0 and node 3)

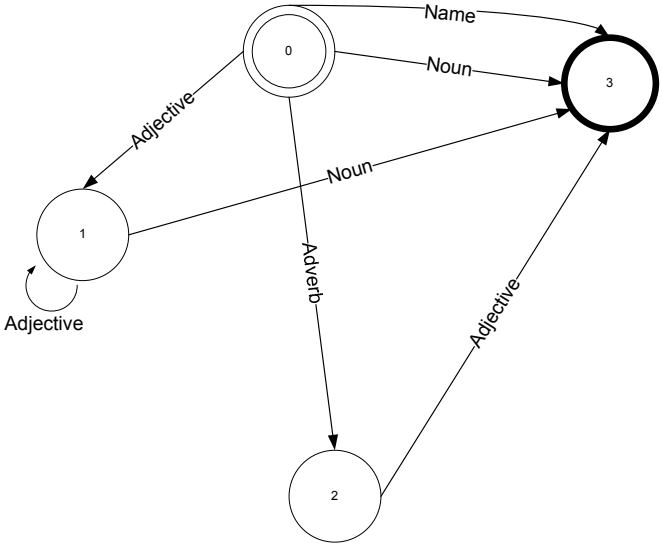


Figure 7. A small segment from the complement part automata

In summary, Table 1 can summarize the parsing stage for the example statement. This result will be the input of the next mapping stage which will need the English-Arabic mapping rule for a "subject, Auxiliary verb, verb, noun". This rule is illustrated in the third row in Table 1.

Table 1. Statement analysis

	Nominal Part	Verbal Part		Complement Part	
S	He	can	can	The	can
English Rule	Subject	Auxiliary verbs	Verb	definite article	Noun
Arabic Rule	Auxiliary Verb	Subject	Verb	The	Noun
TS	يستطيع	هو	تعليب	ال	علبة

In another words, the rule indicates that if "CAN" comes before a verb it will be considered as an auxiliary verb which is different from the meaning of the main verb. This result will affect the translation process and the mapping step. At least, distinguishing between the same words at three different places is imagined to enhance the translation accuracy of about 90%.

Our proposed architecture is to map the sentence into the correct grammatical structure in the target language (i.e., Arabic). If the Transition Network Grammar for the Arabic language is implemented correctly then the applying rules between the two networks should be an easy step. The correct final translation should be " يستطيع هو تعليب العلبة" based on the mapping rule used in this step (see Table 1, row 3 and 4).

5. Conclusion

The proposed architecture in this paper aims to integrate the WordNet ontology with Transition Network Grammar to assist in the parsing procedure within the machine translation process. The proposed architecture can be used between the languages that have real differences in the structure; such as, Arabic and English. It consists of two main stages: syntax analysis and mapping engine. The source text is entered into the syntax analysis stage to be parsed via the Transition Network Grammar guided by the WordNet ontology. The output of this stage will be a full parsed text with clear and correct structure which can be formed as a parsed tree. The second stage addresses the differences between the source and target languages in terms of the structure of the sentences and prefixes and postfixes used to structure the words. The mapping engine is designed to make use of the Transition Network Grammar for the target language to reconstruct the sentences in the correct way after translating the words and phrases in the previous stage. Finally, the usual process in any translation process occurs, that is the actual translation of words and phrases into the target language.

The evaluation for such architecture needs further research in order to design good mapping rules that will work for such purpose. The idea can be limited to cover some kinds of statements as a first stage and then generalized to cover more complex structures. Also, some comparisons should take place to compare in statistics between the proposed architecture and the existing ones in terms of time, accuracy, and complexity.

References

Bataineh B., and Bataine E., 2009, An Efficient Recursive Transition Network Parser for Arabic Language, *in Proceedings of the International MultiConference of Engineers and Computer Scientists*, London, pp.1307-1311

Boretz, Adam, "*AppTek Launches Hybrid Machine Translation Software*" SpeechTechMag.com (posted 2 MAR 2009)

Charniak E., Riesbeck C., McDermott D., and Meehan J., 1987, *Artificial intelligence programming*, Psychology Press.

Elberrichi Z., Rahmoun A., and Bentaalah M., 2008, "Using WordNet for Text Categorization," *The International Arab Journal of Information Technology*, vol. 5, No. 1, pp. 16-24.

Fellbaum C., Hahn U., and Smith B., 2006, "Towards new information resources for public health—From WordNet to MedicalWordNet," *Journal of Biomedical Informatics*, vol. 39, no. 3, June 2006, pp. 321-332.

Goutte C., Cancedda N., Dymetman M., and Foster G, 2009, *Learning Machine Translation*, The MIT Press.

Gruber, T., 1993, "A translation approach to portable ontologies," *Knowledge Acquisition*, 5(2), pp. 199-220.

Guarino N., 1998, "Formal ontology and information systems," *In*: N. Guarino, ed. *Formal Ontology in Information Systems. Amsterdam*, Netherlands: IOS Press, pp. 3-15.

Güvenir H., Cicekli I., 1998, "Learning translation templates from examples," *Information Systems*, Volume 23, Issue 6, Pages 353-363

Hamaydeh, B. and Kharbat, F. (2009) "Phases of English translation engine with partial implementation", BSc graduation project, Zarqa Private University.

Hutchins J., 2003, Machine translation: general overview, chapter 27.

Hwang Y., Finch A., Sasaki Y., 2007, "Improving statistical machine translation using shallow linguistic knowledge," *Computer Speech & Language*, Volume 21, Issue 2, Pages 350-372

Kharbat F., and El-Ghalayini H., 2009, "New Algorithm for Building Ontology from Existing Rules: A Case Study," *in the International Conference on Information Management and Engineering (ICIME 2009)*, April 3-5, Malaysia.

Lopez A., 2008, "Statistical Machine Translation," *ACM Computing Surveys* , 40(3): Article 8, pp. 1-49.

Machine translation, lasted visited june 2010, http://en.wikipedia.org/wiki/Machine_translation

Miller G., "Nouns in WordNet: A Lexical Inheritance System", *International Journal of Lexicography*, vol. 3, no. 4, 1990.

Miller G., 1995, "WordNet: A Lexical Database for English," *Communications of the ACM*, Vol. 38, No. 11, pp. 39-41.

OntologySummit, 2007, WordNet - a networked taxonomy, OntologySummit2007, Population Session, group-B, http://ontolog.cim3.net/file/work/ OntologySummit2007/workshop/Population-spreadsheet_WordNet-as-ontology_20070423b.doc

Placeway, P., 2002, *High-Performance Multi-Pass Unification Parsing.* PhD thesis, Carnegie Mellon University, Pittsburgh, PA. Technical Report CMU-LTI-02-172.

Rodríguez L., García-Varea I., Gámez J., 2008, "On the application of different evolutionary algorithms to the alignment problem in statistical machine translation," *Neurocomputing*, Volume 71, Issues 4-6, Pages 755-765

Salam, K., Khan, M. and Nishino, T., (2008) "Example Based English-Bengali Machine Translation Using WordNet", TriSA 2008, Japan.

Studer R., Benjamins V., and Fensel D. (1998). Knowledge engineering: principles and methods. Data and Knowledge Engineering, 25, pp. 161-197.

Wong F., Dong M., Hu D., 2006, "Machine Translation Using Constraint-Based Synchronous Grammar," *Tsinghua Science & Technology*, Volume 11, Issue 3, Pages 295-306

Faten F. Kharbat is an Assistant Professor in Artificial Intelligence at the Al Ain University for Science and Technology, Abu Dhabi Campus, UAE. She holds PhD degree in Computer Science from the University of the West of England, UK, in 2006. Her main research interest is learning classifier systems, knowledge based systems, applying data mining techniques to business sectors, and recently has been involved in quality of higher education.