December 2006

A Semantic Method to Information Extraction for Decision Support Systems

Bahadorreza Ofoghi
University of Ballarat- Australia

John Yearwood
University of Ballarat- Australia

Ranadhir Ghosh
University of Ballarat- Australia

A Semantic Method to Information Extraction for Decision Support Systems

Bahadorreza Ofoghi
Centre for Informatics and Applied Optimization – University of Ballarat, Australia
bofoghi@students.ballarat.edu.au

John Yearwood
Centre for Informatics and Applied Optimization – University of Ballarat, Australia
j.yearwood@ballarat.edu.au

Ranadhir Ghosh
Centre for Informatics and Applied Optimization
University of Ballarat, Australia
r.ghosh@ballarat.edu.au

ABSTRACT

In this paper, we describe a novel schema for a more semantic text mining process which results in more comprehensive decision making activity by decision support systems via providing more effective and accurate textual information. The utility of two semantic lexical resources; FrameNet and WordNet, in extracting required text snippets from unstructured free texts yields a better and more accurate information extraction process to deliver more precise information either to a DSS or to a decision maker. We explain how the usage of these lexical resources could elevate a focused text mining process which could be applied to an information provider system in a decision support paradigm. The preliminary results obtained after a starter experiment show that the hybrid information extraction schema performs well on some semantic failure situations.

Keywords

Lexical Resources, FrameNet, WordNet, Semantic Text Mining, Information Extraction.

INTRODUCTION

The general definition of decision support systems (DSS), with the specific feature of providing decision makers with the best information while deciding about a problem or situation, covers the information extraction systems which are capable of extracting the most suitable data and information from different resources. As one of the widely used information resources, textual documents are so important containing wide range of information which could be mined and exploited in a decision making process. Since most of such documents are organized in a free unstructured text format, the task of information mining from their content is a crucial activity which should be performed on the basis of both syntactic and semantic background underlying their texts. In this regard, there are semantic barriers in the task of information extraction which interfere and do not allow the system access the intended parts of the text which are indeed the most related snippets to an information need.

To address some of the identified problems in information extraction, pointed in the next sections, we introduce a hybrid schema to overcome syntactic mismatches between semantically related information requests and texts while retrieving answers to an information need submitted to an information extraction system. This semantic approach uses two main lexical resources: FrameNet (Baker, Fillmore et al. 1998) and WordNet (Miller, Beckwith et al. 1990) to better semantic resolution of the texts of an information need and the text of its related documents or passages. The former is used to contribute more structural background human knowledge to texts and make them more informative and structured while the latter is exploited in cases FrameNet fails to add more knowledge to the text; therefore, the WordNet hierarchy can be utilized in finding the most semantically similar concepts and sentences to a given information need.

CHALLENGES IN FOCUSED INFORMATION EXTRACTION

There are many semantic complexities in a free text not well organized in an understandable structure for automated systems which as a result should be capable of text understanding and mining. On the basis of a survey on the question answering

systems, as a type of information systems which articulate information extraction process, participating in the TREC 2004 QA task (Voorhees 2004), we have identified some failure situations in the answer extraction process of such systems which interfere and reduce the accuracy of answering the questions:

- Ignorance or unawareness of scenario-based relations between language items; this happens, for instance, in a question like "*Who is his mother?*" where in the text of the passages "*he is the son of X*" has been mentioned.
- Far or wrong pronominal anaphora referencing which makes the task of the justification of a scenario hard while referencing to the explicitly mentioned entities in text.
- Inability in accessing indirectly available objects hidden behind a chain of lexical relations; as an instance of this situation, consider the question "*Who discovered quarks?*" and the answer passage "*In 1974, using beams of electrons and antielectrons, or positrons, Richter discovered particle that came to be called Psi/J. It contained two quarks possessing a previously unknown flavor called charm*". After finding different relations there is a chain as below which makes the answering process too complicated with respect to semantics behind the chain:

  Richter *discovered* ***particle*** *called* ***Psi/J*** *contained* ***quarks***

- Failing in resolution of deep unstructured semantic relatedness between terms when there is only background semantic knowledge to resolve similarity issues. To express the depth of the semantics required and to show the difficulty of resolving these situations, suppose the question "*When did he start writing?*" and the passage "*Joe put pen to paper in the ninety's*". The main clue to finding any relation between the passage and the question in this example is the background knowledge of the concept "*writing*" and its relation with two other concepts "*pen*" and "*paper*". From the syntax point of view, this is a projection of a verb into two nouns and another verb "*put*" which does not have any relation with the verb "*write*" without considering the contextual nouns "*pen*" and "*paper*".

The main top level reason to explain why these malfunctions happen seems to be the nature of Named Entity Oriented information extraction approaches used by most question answering systems. Whilst in many cases, the correct answer to a question is either not of a solid Named Entity type or contains more than a unique Named Entity term.

On the other hand, regarding the failure situations again, it is noticeable that such conditions are related with some well-known problems in computational linguistics like as Anaphora Reference Resolution (Lappin and Leass 1994), Lexical Chain Resolution (J. and G 1991), and Semantic Role Labeling (Gildea and Jurafsky 2002) (Litkowski 2004).

As a result of these analyses, we are convinced that the contribution of the extra knowledge by semantic lexical resources, from computational linguistics point of view, to texts could overcome most occasions in which syntactic mismatches interfere in information extraction.

SEMANTIC LEXICAL RESOURCES

As already mentioned, two major English lexical resources are used in our information extraction schema to semantic resolution of the texts of the information requests and their most related documents or passages: WordNet and FrameNet, as introduced in the next sections.

WordNet as a Concept Hierarchy

WordNet is a lexical reference system whose design is inspired by psycholinguistic theories of human lexical memory (Miller, Beckwith et al. 1990). This system includes all English verbs, nouns, and adjectives organized into synonym sets. There are different relations between the synonym sets, also known as synsets. Each of the sets represents an underlying concept and from this point of view, this lexical system forms a concept hierarchy with different abstraction levels for concepts. The main organization of WordNet consists of the semantic relations between synsets, as a semantic relation is a relation between meanings and the meanings are expressed in synsets. The semantic relation set in WordNet contains the relations Synonymy, Antonymy, Hyponymy, and Meronymy. There are also Morphological Relations between word forms to deal with inflectional morphology in the language (Miller, Beckwith et al. 1990).

FrameNet as a Concept-Scenario Network

FrameNet is a lexical resource for English (Baker, Fillmore et al. 1998) that relies on an infrastructure based on Frame Semantics (Fillmore 1976) (Lowe, Baker et al. 1997). It should be noticed that Frame Semantics is different from Marvin Minsky's frames (Minsky 1974). Frame Semantics tries to emphasize the continuities between language and human experience, and FrameNet as a result of these efforts, is a framework to encapsulate the semantic information gained via Frame Semantics (Petruck 1996).

Generally, the main entity in FrameNet is a Frame as a kind of generalization over concepts semantically related to each other. The semantic relation between concepts in a Frame is realized with regard to the scenario of a real situation which may happen and cover the participant concepts rather than synonymy or other dictionary-oriented peer-to-peer relations. In this regard, Frames encode the base definitions necessary to understand the semantics and the scene of each contained term. In other words, real-world knowledge about real scenarios and their related properties are encoded inside the Frames (Lowe, Baker et al. 1997). To address such a feature, each Frame contains some Frame Elements as representatives of different semantic and syntactic roles (also known as valences) regarding a target word inside the Frame. The semantic roles are usually common properties among all of the terms that are inherited from a Frame. This ensures a suitable inclusion over the English terms which either have similar meanings or share the context and/or the scenario in which they could occur in the sentences of the language.

Since there are semantic relations between the circumstances covered by different Frames, a limited set of Frame-to-Frame relations has been defined in FrameNet which connect Frames to constitute a network of concepts and their pictures (Ruppenhofer, Ellsworth et al. 2005). The current set of Frame-to-Frame relations (i.e. Subframe, Inheritance, SeeAlso, Using, Inchoative-of, and Causative-of) has been designed to capture the FrameNet I concept of Domains where Frames were organized by domains; very general categories of human experience and knowledge to provide useful groupings of semantic Frames (Ruppenhofer, Ellsworth et al. 2005).

FrameNet is different from WordNet as it contains not only words with similar meanings, but also higher level concepts of similar scenarios of usage in the real-world. On the other hand, these scenarios are related to each other to model an end-to-end scenario containing some smaller sub-scenarios. The different relation types existing between Frames cover this overview of the different events all of which could be realized by FrameNet.

A HYBRID INFORMATION EXTRACTION SCHEMA

To address semantic barriers resulting in reduced capability in extracting focused text snippets to information needs a hybrid information extraction schema is introduced which benefits from the above mentioned semantic lexical resources. An abstract information system may contain three major sections: i) the Information System's Interface to communicate with the user individual or system evoking the information system, ii) the Information System's Engine to perform main activities of the system with respect to the requirements the system is supposed to satisfy, and iii) the Arbitrary Interface to play the role of the bridge between the system and any type of information resources, other agents, higher or lower level communicating systems and so on.

In such a general view, we are interested in the Information System's Engine where the task of Information Extraction is performed when the system needs to be exploited in a decision support paradigm. The information extraction task can operate either on a list of related documents or in an ordered list of smaller text passages which are more precisely and in a more focused fashion related to the information need and its answer category. The hybrid schema for information extraction, as depicted in Figure 1, includes: i) Information Extraction Engine, ii) a Semantic Interface, and iii) two Semantic Lexical Resources.

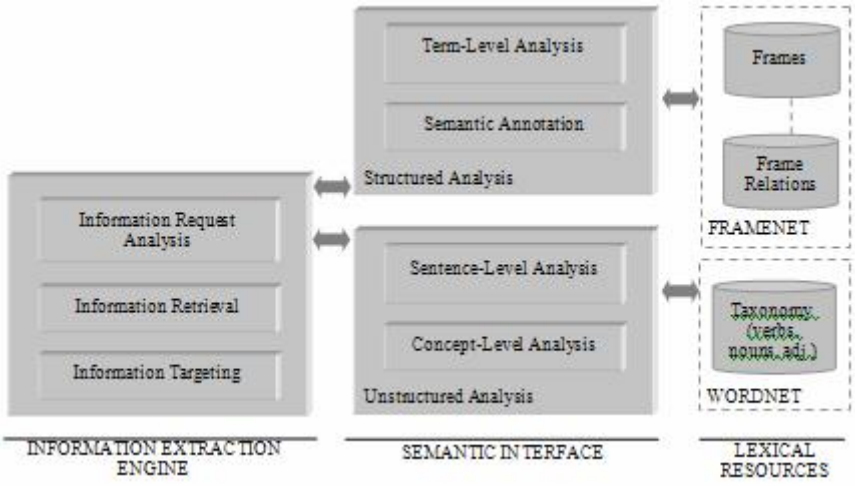

Figure 1. The Hybrid Information Extraction System

The Information Extraction Engine firstly analyzes the information request to the extent that what its answer category is and what its main keywords are. Having found about these, the second task of this interface is to retrieve and rank the most related textual documents and then passages with the information request. The task of information retrieval could be achieved by a text retrieval system which uses classic document and passage indexing and representation. The third activity of the Information Extraction Engine is to extract the most related pieces of information from the content of the ordered list of documents and/or passages already retrieved. The Information Targeting module is responsible to activate the appropriate Text Analysis process via Semantic Interface in order to get enough background knowledge and extract the pieces of information requested.

The Semantic Interface, which is responsible for providing this engine with more lexical semantic information, performs three tasks. Firstly, it is capable of annotating the text with the Frames and their corresponding Frame Elements of FrameNet when there are Frames evoked by the elements of the text. Secondly, the interface can do term-level analysis, having known the Frame already invoked by a term. Thirdly, this interface is able to measure semantic relatedness between two sentences with regard to the taxonomy of WordNet, whenever it is necessary.

The main task of information extraction using the semantic interface in our schema is a hybrid process consisting of two kinds of text analysis: i) structured analysis, and ii) unstructured analysis. In cases that the text of an information need evokes at least a single Frame of FrameNet and there is at least a match Frame evoked by the text of the related documents or passages retrieved by the information retrieval engine, the structured analysis is performed in order to identify the vacant Frame Element of the invoked Frame of the information need and find the match Frame Element in the related texts and report its instance value as the candidate piece of information. In such cases, the information inside a single Frame is of importance. We refer to this term-level related analysis as *Intra-Frame Analysis* to underpin the scenario-based relations between syntactically different situations in texts. For example, if there is an information need like "*report the name of Napoleon's son*" and there is a passage containing the sentence "*Napoleon was Julian's father …*", such analysis empowers the system to report "Julian" as a candidate piece of information, although the texts of the passage and information need drastically differ from each other in syntax. The information retrieval task of such mismatched syntaxes is a complex task which requires semantic information be articulated by the retrieval process (Ofoghi, Yearwood et al. 2006). The nature of the encapsulated Frame Semantics in FrameNet allows the information system to cope with word mismatches with the same meaning (e.g. discover and spot), word mismatches with the same scenario coverage (e.g. son and father), parts-of-speech mismatches with the same meaning (e.g. beater and beat), and role mismatches in the same scenario (e.g. sender and receiver).

The structured analysis, although sophisticated, is not able to underpin the elements of texts when there is no Frame evoked further to not coverage of the sentence scenario by FrameNet. Also, there is always a possibility that the same Frame as evoked by an information need is not invoked by the text of the related documents or passages. Due to such conditions, the unstructured analysis is activated to report less confident answer pieces. This analysis starts with measuring the semantic relatedness of each sentence of related texts with the sentence of the information need. In order to calculate such relatedness, the WordNet hierarchy is exploited. To have the similarity between sentences measured, it is necessary to break the sentences down into their elements (i.e. phrases and words) and measure the semantic similarity between them. The sentence-to-sentence relatedness algorithm that we have used was developed by Troy Simpson and Thanh Dao[1] which benefited from WordNet-based Path Length similarity measures between concepts.

Having all of the sentences scored with their semantic similarity to the question, they are ranked and all Named Entities from inside these sentences are extracted. The set of Named Entities is further processed to be filtered against the category of the information need already identified by Information Need Processing module. After extracting all Named Entities and filtering them with respect to the category of the information need, any remaining Named Entities are scored and ranked and finally, the top most Named Entities in the ranked list are reported as candidate pieces of information.

The unstructured analysis is a step towards resolving deep syntactic mismatches and semantic relations between not only words and phrases but the sentences of language. Suppose the information need like "*tell me when Mike started writing*" and the potential sentence as "*Mike put pen to paper when he was 40, in 1945*". In this case, there is no similar or the same predicate between the information need and the answer sentence, though the sentence contains the exact piece of information needed. We refer to these situations as deep unstructured syntactic mismatches between semantically related texts.

Table 1 summarizes the two main semantic barriers, out of four mentioned in section 2, that we have identified and how they are resolved using the hybrid information extraction schema.

[1] http://www.codeproject.com/csharp/semanticsimilaritywordnet.asp

Semantic barrier	Our model's analysis	Example
Scenario-based relation	Structured analysis	son – father beat – knock down …
Deep unstructured semantic relation	Unstructured analysis	write – put pen to paper …

Table 1. Main problems explored and resolved by our hybrid information extraction schema

PILOT EVALUATION

Question Answering (QA) systems, as one of the extreme types of information systems, could be realized as: i) a stand-alone DSS system which interactively provide decision makers with required information, and ii) an automated information provider system as part of an individual DSS system. Figure 2 depicts these two views. The left picture shows that the output of the question answering system is the final result submitting to the end-user, while the picture at right shows a DSS paradigm in which the question answering system only plays the role of the Data Retriever to a more general information system which then models the output of the question answering system to make it more suitable for decision makers with regard to the context of the problem.

Regardless of what categorical point of *view* such systems are be perceived from, their performance in answer extraction, which is a hidden information extraction sub-module, is so crucial and important that affects the total end-to-end accuracy of the system resulting in varying user judgments and conclusions.

With respect to this, we have developed a semi-automated question answering system to test the performance of the information extraction module implemented on the basis of the proposed hybrid schema. Preliminary results, compared with the outstanding systems participating in TREC 2004 (Voorhees 2004), show that the hybrid model is performing well to overcome failure situations of other systems.

Questions	Total number	Percentage correctly answered by our system	Our system's MRR	Baseline MRR
Correctly answered by baseline system	5	100%	0.4733	1.0000
Correctly unanswered by baseline system	10	80%	0.6833	0.0000
All subset	15	86.66%	0.6133	0.3333

Table 2. Results obtained for a subset of TREC 2004 factoid track

We have used TREC 2004 QA Track and its related text collection called AQUAINT. As the standard evaluation and test bed for QA systems, TREC provides a set of fact-seeking (factoid) questions whose answers could be identified from AQUAINT. The answering process of such questions is highly dependent to the information extraction mechanisms as the answer extraction module where participant systems search related pieces of texts to extract a focused and succinct text snippet in respond to a given factoid question. We ran the system for one subset of the track containing 2.13% of all factoid questions (5 questions) answered by the baseline system and 4.26% of all factoid questions (10 questions) not correctly answered by this system (totally 15 factoid questions out of 230). The questions were selected randomly just with regard to the result of the baseline system to categorize them into correctly answered and unanswered questions to evaluate the proposed schema with respect to the already resolved challenges by other systems as well as the incorrectly answered questions where existing systems fail to find correct pieces of information. The main evaluation criteria in TREC competitions is Mean Reciprocal

Rank abbreviated as MRR[2]. Our results show that 100% of the first category (MRR=0.4733) and 80% of the second category (MRR=0.6833) have been answered by our question answering system which benefits from the hybrid information extraction mechanism as its core text processing module. Table 2 summarizes the results.

The results, obtained on the basis of a preliminary study, show that not only does the proposed methodology of semantic resolution work on the already solved situations by existing systems, but also it is capable of making some progress (MRR=0.6833) on top of the structured and unstructured semantic situations where the baseline QA system fails (MRR=0.0000). The last row of the table, consequently, depicts how the proposed system has improved the overall MRR over the 15 randomly selected questions from 0.3333 by the baseline system to 0.6133.

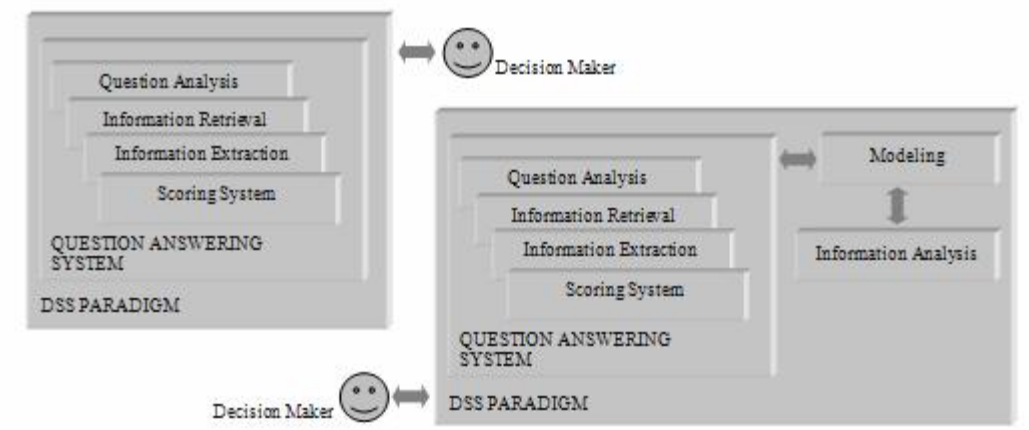

Figure 2. Two Views of Question Answering Systems and DSS

The prototype system to get the results is a semi-automated system implemented and integrated under Windows OS (in Visual C#.Net). The main difficulty to make such a system operate fully automated is the annotation task of the free text of the passages (or documents) related to the question. For the time being, there is no software package to make the role labeling automated, to our knowledge. In our experiments, we manually annotated the passages retrieved by our passage retrieval system (~150 passages, i.e. 10 passages per question) against the Frames and Frame Elements of FrameNet[3].

CONCLUSION

To improve the effectiveness of information extraction in the context of DSS systems, we have developed a hybrid schema to deliver more accurate and semantic identification and extraction of pieces of information from free texts. This approach requires the two well-know semantic lexical resources for English; FrameNet and WordNet, to cope with both structured and unstructured semantic situations in texts while trying to answer focused information needs. Some computational linguistic tools (e.g. anaphora reference revolver) are necessary to be used in such an information system. We are convinced, on the basis of some preliminary results, that the utilization of semantic lexical resources to contribute more human knowledge to free texts could drastically elevate the performance of information extraction systems articulated in a knowledge-driven decision making paradigm.

FUTURE WORK

We will be trying to justify our preliminary work on bigger subsets (up to the whole set) of factoid questions provided by TREC to make more valid and comparable arguments with respect to the baseline QA systems. At the same time, we will be to find a semantic approach to cope with the lexical chains in order to make the system capable of finding answers behind a chain of lexicalized semantic relations between a number of language items.

[2] http://trec.nist.gov/data/qa.html
[3] http://framenet.icsi.berkeley.edu/

REFERENCES

1. Baker, C. F., C. J. Fillmore and J. B. Lowe (1998). The Berkeley FrameNet Project. *International Conference On Computational Linguistics* 1: 86 - 90.
2. Fillmore, C. J. (1976). Frame Semantics and the Nature of Language. *In Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech* 280: 20-32.
3. Gildea, D. and D. Jurafsky (2002). Automatic Labeling of Semantic Roles. *Computational Linguistics* 28(3): 245 - 288.
4. J., M. and H. G (1991). Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. *Computational Linguistics* 17(1): 21-42.
5. Lappin, S. and H. J. Leass (1994). An Algoritrhm for Pronominal Anaphora Resolution. *Computational Linguistics* 20(4): 535-561.
6. Litkowski, K. (2004). Senseval-3 Task: Automatic Labeling of Semantic Roles. *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*: 9-12.
7. Lowe, J. B., C. F. Baker and C. J. Fillmore (1997). A Frame-Semantic Approach to Semantic Annotation. *SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*
8. Miller, G., R. Beckwith, C. Fellbaum, D. Gross and K. J. Miller (1990). Introduction to WordNet: an On-Line Lexical Database. *International Journal of Lexicography*: 235-244.
9. Minsky, M. (1974). A Framework for Representing Knowledge. *The Psychology of Computer Vision*: 211-277.
10. Ofoghi, B., J. Yearwood and R. Ghosh (2006). A Semantic Approach to Boost Passage Retrieval in Question Answering. *The 29th Australian Computer Science Conference* 48: 95-101.
11. Petruck, M. R. L. (1996). Frame Semantics. *Handbook of Pragmatics Online*.
12. Ruppenhofer, J., M. Ellsworth, M. R. L. Petruck and C. R. Johnson (2005). FrameNet: Theory and Practice. *http://framenet.icsi.berkeley.edu/*.
13. Voorhees, E. M. (2004). Overview of the TREC 2004 Question Answering Track. *The Thirteenth Text REtrieval Conference Proceedings (TREC 2004)*.