International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:8, No:8, 2014

A Web-Based Self-Learning Grammar for Spoken Language Understanding

S. M. Biondi, V. Catania, R. Di Natale, A. R. Intilisano, D. Panno

***Abstract*—**One of the major goals of Spoken Dialog Systems (SDS) is to understand what the user utters.

In the SDS domain, the Spoken Language Understanding (SLU) Module classifies user utterances by means of a pre-definite conceptual knowledge. The SLU module is able to recognize only the meaning previously included in its knowledge base. Due the vastity of that knowledge, the information storing is a very expensive process.

Updating and managing the knowledge base are time-consuming and error-prone processes because of the rapidly growing number of entities like proper nouns and domain-specific nouns. This paper proposes a solution to the problem of Name Entity Recognition (NER) applied to a SDS domain. The proposed solution attempts to automatically recognize the meaning associated with an utterance by using the PANKOW (Pattern based Annotation through Knowledge On the Web) method at runtime.

The method being proposed extracts information from the Web to increase the SLU knowledge module and reduces the development effort. In particular, the Google Search Engine is used to extract information from the Facebook social network.

***Keywords*—**Spoken Dialog System, Spoken Language Understanding, Web Semantic, Name Entity Recognition.

I. Introduction

A Spoken Language Understanding (SLU) module is used in a Spoken Dialog System (SDS) domain to understand the meaning of the utterances of a user. Most SLU modules are based on a Context Free Grammar (CFG) that provides the definition of the meaning of each word and the set of utterances related to it. This definition requires one or more grammar files containing a set of words and their meanings. This approach generates a SDS for a given application by using a specific CFG.

One of the major obstacles in the development of a spoken dialog application is the writing of a domain-specific grammar. The development of a grammar is a very complex task when the used language is inflective [1], [2].

The grammar is updated in a semi-automatic and unsupervised way by the automatic retrieving of the relations between meaning and user utterances.

The proposed method is able to discover a relationship between an expression of the user and its meaning and to update the grammar in a semi-automatic and unsupervised way.

The web represents a large corpus of information which can be used to extract knowledge and automate the association of a word with its meaning. In this context, the web can be considered as corpus that is continuously updated and from which it is possible to retrieve knowledge to define a grammar in a semi-static way.

We introduce an unsupervised approach for the SDS domain based on the PANKOW method [3] (Pattern based Annotation through knowledge on the Web) to use knowledge from the web to update the CFG in grammar files. Thanks to this method an SDS can recognize utterances not included in the grammar initially. Our technique uses these words to associate words with their respective meanings. The PANKOW method uses the Google Search Engine to retrieve information from the web and it requires a list of meanings that the Google Search Engine uses in its queries. In particular, the choice of the pair word-meaning will be made according to the confidence level of the relations derived from the Google Search Engine results.

Furthermore, the system can retrieve information from Facebook through Google to increase its list of *names* or *surnames*.

This paper is organized as follows: a list of related works is presented in Section II; the PANKOW method in the SDS domain is described in Section III; the unsupervised self-learning grammar method is described in Section IV; Section V shows the results of a set of tests; finally, the conclusions are presented in Section VI.

II. Related Work

For the SDS domain, the problem described above has been addressed in [4]. In particular, that paper shows how to implement an incremental On2L ontology to obtain knowledge at runtime. That system for *semantic annotation* is based on models of the natural language to extract the relations between a term and its corresponding concept in the ontology. All unrecognized words are classified as out-of-vocabulary (OOV).

That paper shows the advantage of making a distinction between global OOVs and local OOVs. Global OOVs include, for example, proper nouns of stars, movies, hotels, while local OOVs include proper nouns or names of countries or cities. In the first case, a search on Wikipedia is performed; in the second case, the Google Search Engine is used. To be more precise, if the OOV typology is unknown, the system performs a search on Wikipedia as a first step, and then it uses Google Search. If the search on Wikipedia produces results, these are used as a test for Google Search; otherwise, a search with

S. Biondi, V. Catania, R. Di Natale, A.R. Intilisano, and D. Panno are with the Dipartimento di Ingegneria Elettrica Elettronica ed Informatica, Università di Catania, Catania, Italy (e-mail: salvo.biondi@dieei.unict.it, vincenzo.catania@dieei.unict.it,raffaele.dinatale@dieei.unict.it, aintilis@dieei.unict.it, daniela.panno@dieei.unict.it, respectively).

Google Engine is performed. However, that should be considered as a work in progress and it has not been evaluated in detail so far. Its aim was to provide a task-oriented evaluation setup and to improve the results by using different techniques.

Different fields exploit semantic annotation to retrieve information from web. For example, it is used in Biomedical Literature to implement an automatic extraction of biomedical information [5].

III. Pankow Method in SDS Domain Application

The web contains millions of documents, forming a big corpus that can be used as a source of information in the development of an auto-updating grammar.

The core of this application is based on PANKOW, a self-annotating method that uses linguistic patterns to annotate the meaning of a word by means of the Web. The PANKOW method is an unsupervised pattern-based approach that categorizes instances according to a series of concepts.

The self-annotating method is based on a series of Linguistic Patterns such as the HERST pattern [6] that creates regular expressions in order to identify instance-concept relations in a text.

In this section we describe the major aspects of the PANKOW method and the way it has been adapted to our application.

The PANKOW method uses a list of proper nouns as inputs. Every word uttered by the user is analyzed by the self-annotating method at runtime. Therefore, the user utterance represents the new input to the PANKOW method. Furthermore, the PANKOW method initially requires a series of concepts along with a series of linguistic patterns to perform self-annotation using the Web corpus. These concepts, in our domain, represent the meanings that the words can assume.

The second step consists in deriving phrases by introducing an instance (word) and ontological concepts (meaning) potentially related to it in the linguistic patterns. For example:

<INSTANCE|WORD> IS A <CONCEPT | MEANING>

If the user utters the word “Italy” and “country” and “hotel” are the candidate meanings, the phrases derived by using the linguistic pattern “IS A” are: “Italy is a country” and “Italy is a hotel”.

The second step ends when all possible phrases are derived, and these become the inputs for the third step, in which these are used as queries to the Google Search Engine. The Google Engine is called through its API and the retrieved results represent the number of hits for each phrase. Then, according to the results obtained, the service selects the meaning (among those in the list) to be associated with the utterance provided by the user. The following formula allows to retrieve the number of occurrences of an instance *i* of a concept *c* for all patterns *p* belonging to the set *P*.

$$count(i,c) = \sum_{p \in P} f(i,c,p)$$

In the literature, there are several approaches for analyzing a text corpus. They capture specific information or relations between words and concepts.

HEARST describes an automatic acquisition method of hyponymy lexical relations from unrestricted text. HEARST identifies a set of lexical-syntactic patterns that occur frequently in a text corpus. These may in fact indicate lexical relations that could be of interest.

Linguistic patterns [3] are formulated using the variable ‘<INSTANCE>’ to refer to a candidate noun phrase, such as the name of an ontologic instance, while the variable ‘<CONCEPT>’ refers to the name of a given concept. The HEARST patterns used in the PANKOW method are the following:

<CONCEPT>s such as <INSTANCE>
such <CONCEPT>s as <INSTANCE>
<CONCEPT>s, (especially|including)<INSTANCE>
<INSTANCE> (and|or) other <CONCEPT>s

As reported in [3], the verb "to be" defines an entity as an instance of a concept. A possible linguistic pattern is:

<INSTANCE> is a <CONCEPT>

Our method uses these patterns to realize an automatic unsupervised approach for discovering lexical relations between the utterance of a user and its meaning using the Web.

IV. Unsupervised Self Learning Grammar Method Using Olympus Framework

The self-learning grammar system being proposed uses a revised version of the PANKOW method that is used to improve the performance of a SLU module.

Olympus represents the state-of-art framework to create a SDS [7]. Olympus is a complete framework for implementing SDSs created at Carnegie Mellon University (CMU). Olympus includes a dialog manager called RavenClaw [8], which supports mixed-initiative interaction, as well as SLU components that handle speech recognition (Sphinx), understanding (Phoenix) and generation (Rosetta).

In particular, Phoenix [9] is an easy and robust Natural Language Processing (NLP) system that performs the human language syntactic analysis based on a grammar. A specific Spoken Language Application requires the development a context-free grammar. According to the described grammar, the SDS system can understand the utterances pronounced by a user.

The Phoenix grammar file definition comes in the form of a slot definition. A slot is a symbol that captures a subset of the language, such as "place names" or "country" of “surname”. The general structure of a slot definition is as follows:

```
[Country]
    (Italy)
    (South Africa)
    (Congo)
    (Guatemala)
;
```

The goal of the system is to simplify the grammar development process, automatically detecting the relations between utterances and meanings and it does not require pre-definite knowledge of the meaning of a word. It works by merging the Phoenix parser with the PANKOW service. Therefore, the slot definition changes according to the service features.

For example, the grammar slot “country”, represents an ontological concept candidate of PANKOW, and it is also the meaning that the system tries to associate to the user utterances “Italy” and “South Africa”.

If the user utters “United States”, the SLU parser will not recognize this word as a "country" meaning, because it is not included in the grammar slot definition. In fact, Phoenix does not recognize words not included in the grammar slot definition. Including all possible words for each concept is a very expensive process.

Our system, thanks to the Unsupervised Self learning grammar method, only requires the definition of the symbol of the grammar slot like “country”, and it represents the meaning of a specific list of utterances. The meaning of the word "country" is retrieved from the web using the PANKOW method. This method attempts to discover if a linguistic relation exists between “country” (meaning) and a specific word ("United State"). The new grammar slot definition is shown below:

```
[Country]
    (Italy)
    (South Africa)
    (Congo)
    (Guatemala)
   Africa           (auto-generated by our method)
   United States (auto-generated by our method)
   Australia       (auto-generated by our method)
;
```

The PANKOW method updates the old grammar file, adding the instances found.

When the Phoenix parser does not recognize a word, it will become a candidate for our unsupervised method. The scheme of the described system is shown in Fig. 1.

A series of tests have been performed to evaluate the proposed approach. The input for these tests was a list of items whose meaning is known. Such tests show that the unsupervised method without pre-defined CFG is able to retrieve large amounts of information from the web.

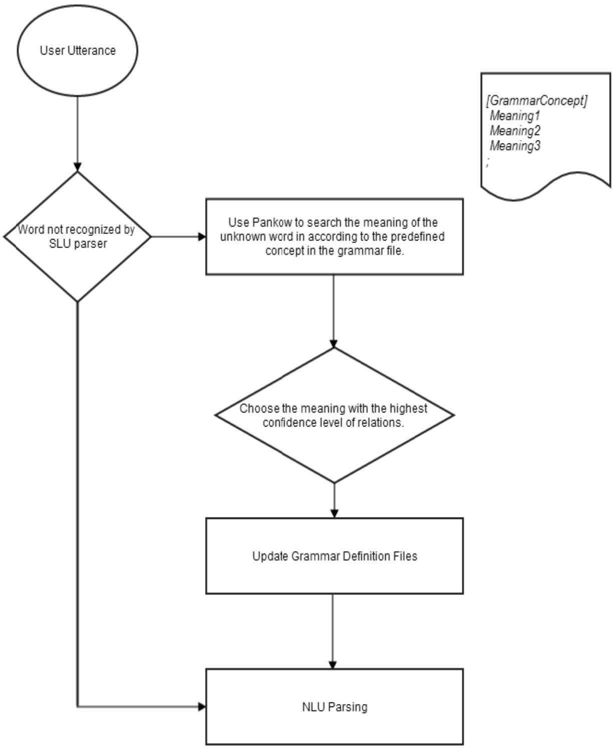
User Utterance
[GrammarConcept]
Meaning1
Meaning2
Meaning3
;
Word not recognized by SLU parser
Use Pankow to search the meaning of the unknown word in according to the predefined concept in the grammar file.
Choose the meaning with the highest confidence level of relations.
Update Grammar Definition Files
NLU Parsing

Fig. 1 A self-learning Grammar System Schema

V. Test and Result

This section shows the result of the performed tests. Initially a grammar file was created as defined by “new grammar slot definition”, with the choice of a series of Concepts.

A list of random items is the input to the unsupervised method.

Below, we show the concepts used in our method for the tests:

Concept (Meanings)

- Name
- Surname
- Smartphone
- Country

The input files contain a list of (Name, Surname, Smartphone and Country). For each input term, the unsupervised method tries to find a linguistic relation between the input item and all concepts. Thus, the Google Search Engine returns the number of the f(c,i,p) (Concept, Instance, Pattern) occurrences.

The number of occurrences generated by Google Search determines the confidence level of the relations between an instance and a concept, as shown in Table I. The results show a high percentage of hits of concepts like “country” and “smartphone”. Although name and surname show smaller values, their percentages of hits are still relatively high. “Name” and “Surname” represent concepts with a larger probability of generating ambiguities than “smartphone” or “country”, as these are generally more specific terms.

TABLE I
RESULT TEST UNSUPERVISED METHOD

Meaning	Number of instance	Hit percentage
Name	50	58%
Surname	50	60%
Country	50	94%
Smartphone	50	94%

We introduce a service to retrieve information about the concepts “name” and “surname”. Since these concepts are strictly related to one or more individuals, social networks such as Facebook may constitute an important source of information.

If a user queries Google Search with the name and surname of an individual, the first results are usually related to Facebook. Therefore, our method questions the search engine with queries like “Name Facebook” or “Surname Facebook”. This is used verify the existence of an individual that possesses a certain name or surname. Usually an individual's Facebook profile contains first name, other name and Surname. The method captures the meaning of the words, relying on their positions. In particular, the system uses the first 10 results of the Google Search Engine. The results are shown in Table II.

TABLE II
RESULT TEST UNSUPERVISED METHOD

Meaning	Number of instance	Hit percentage
Name	50	44%
Surname	50	88%

VI. CONCLUSIONS

The system we propose generates a grammar automatically based on information retrieved from the web. The obtained results show that the system is able to extract a significant amount of knowledge from the web without human assistance and without predefined knowledge, thus reducing the writing a grammar considerably. This approach combines manual generation of a grammar with the auto-updating of the same grammar. Information from Facebook could be used with good results to realize NER in particular for surnames (88%). Thanks to this system, the developing complexity of CFG for a SLU decreases significantly.

This System could be enhanced with the use of auto-adaptive linguistic patterns based on the context, or using directly the Facebook API to retrieve information about names and surname.

ACKNOWLEDGMENT

The authors were supported by the Sicily Region grants PROGETTO POR 4.1.1.1: "Rammar Sistema Cibernetico programmabile d’interfacce a interazione verbale".

REFERENCES

[1] V. Catania, R. Di Natale, A. R. Intilisano, Y. Cilano, D. Panno. "SmartGrammar: A dynamic spoken language understanding grammar for inflective languages".(In press)

[2] S. M. Biondi, V. Catania, Y. Cilano, R. Di Natale and A.R. Intilisano. 2014. An Easy and Efficient Grammar Generator for Spoken Language Understanding, The Sixth International Conference on Creative Content Technologies – Vol. 7 nr 1 and 2, Venice, Italy.

[3] Philipp Cimiano, Siegfried Handschuh, and Steffen Staab. 2004. Towards the self-annotating web. In *Proceedings of the 13th international conference on World Wide Web* (WWW '04). ACM, New York, NY, USA, 462-471.

[4] Berenike loos. 2006. On2L - A Framework for Incremental Ontology Learning in Spoken Dialog Systems. In *Proceedings of the COLING/ACL 2006 Student Research Workshop*, Sydney, Australia.

[5] Rune Sætre, Amund Tveit, Tonje S. Steigedal, and Astrid Lægreid. 2005. Semantic annotation of biomedical literature using google. In *Proceedings of the 2005 international conference on Computational Science and Its Applications - Volume Part III* (ICCSA'05), Osvaldo Gervasi, Marina L. Gavrilova, Vipin Kumar, Antonio Laganà, and Heow Pueh Lee (Eds.), Vol. Part III. Springer-Verlag, Berlin, Heidelberg, 327-337.

[6] Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics - Volume 2* (COLING '92), Vol. 2. Association for Computational Linguistics, Stroudsburg, PA, USA, 539-545.

[7] Dan Bohus, Alexander I. Rudnicky, “The RavenClaw dialog management framework: Architecture and systems”, _Computer Speech and Language, _vol. 23, no. 3, 2009

[8] Bohus, A. Raux, T. K. Harris, M. Eskenazi, and A. I. Rudnicky, “Olympus: an open-source framework for conversational spoken language interface research,” in *proceedings of HLT-NAACL 2007 workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technology*, 2007.

[9] W. Ward, "Understanding spontaneous speech: the Phoenix system," *Acoustics, Speech, and Signal Processing, 1991. ICASSP- 91, 1991 International Conference on,* 14-17 Apr 1991, pp.365-367 vol.1.