POPULATING A DOMAIN ONTOLOGY FROM A WEB BIOGRAPHICAL DICTIONARY OF MUSIC

An Unsupervised Rule-based Method to Handle Brazilian Portuguese Texts

Eduardo Motta, Sean Siqueira and Alexandre Andreatta
Department of Applied Informatics, Federal University of the State of Rio de Janeiro (UNIRIO)
Av. Pasteur, 458, Urca, Rio de Janeiro, Brazil

Keywords: Information extraction, Ontology population, Natural language processing, Brazilian Portuguese.

Abstract: An increasing amount of information is available on the web and usually is expressed as text, representing unstructured or semi-structured data. Semantic information is implicit in these texts, since they are mainly intended for human consumption and interpretation. Since unstructured information is not easily handled automatically, an information extraction process has to be used to identify concepts and establish relations among them. Information extraction outcome can be represented as a domain ontology. Ontologies are an appropriate way to represent structured knowledge bases, enabling sharing, reuse and inference. In this paper, an information extraction process is used for populating a domain ontology. It targets Brazilian Portuguese texts from a biographical dictionary of music, which requires specific tools due to some language unique aspects. An unsupervised rule-based method is proposed. Through this process, latent concepts and relations expressed in natural language can be extracted and represented as an ontology, allowing new uses and visualizations of the content, such as semantically browsing and inferring new knowledge.

1 INTRODUCTION

This paper describes an implemented method to extract information from Portuguese texts on a rule-based unsupervised manner. The information extraction process output is represented as a domain ontology.

An increasing amount of information is available on the web and is frequently expressed as text, representing semi-structured or unstructured data. However, computational processes don't easily handle unstructured information. Information extraction (IE) seeks to structure information by adding meaning to raw data. It is defined by Moens (2006) as "the identification, and consequent or concurrent classification and structuring into semantic classes, of specific information found in unstructured data sources, such as natural language text, making the information more suitable for information processing tasks."

A usual way to capture and represent knowledge of a specific field (a knowledge domain) is through the use of a domain ontology. According to Gruber (2008), "an ontology defines a set of representational primitives with which to model a domain of knowledge or discourse". The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members).

Therefore, the outcome of an IE process can be represented as a domain ontology. Ontologies are an appropriate way to represent structured knowledge bases, enabling sharing, reuse and inference.

In this work, we present the implementation of an IE process that populates a domain ontology. The knowledge domain handled by our solution is the biographical and artistic data on the music field.

The source of information is a biographical dictionary of Brazilian popular music and the output of the process is a populated domain ontology.

Through this process, latent concepts and relationships expressed in natural language can be extracted and represented, allowing new uses of the available content, like integrating with other information sources (based on semantic concepts) or navigating the knowledge base on a hierarchical manner. Moreover, after knowledge is represented as a populated domain ontology that is enriched with

Motta E., Siqueira S. and Andreatta A.
POPULATING A DOMAIN ONTOLOGY FROM A WEB BIOGRAPHICAL DICTIONARY OF MUSIC - An Unsupervised Rule-based Method to Handle Brazilian Portuguese Texts.
DOI: 10.5220/0001842301920199
In *Proceedings of the Fifth International Conference on Web Information Systems and Technologies (WEBIST 2009)*, page
ISBN: 978-989-8111-81-4

logical rules, inference engines can be used to derive new knowledge not explicitly contained in the source texts.

One of the benefits of representing the textual content as a structured knowledge base is to allow browsing through dimensions like year or place of birth, type of instruments performed or even navigating through relationships among artists, like partnership.

The approach of this work is to exploit information that is most relevant to the domain, such as biographical data (e.g., date and place of birth and death, and genealogical relations), artistic data (e.g., partnership) and career events (like recording or releasing a CD, composing a song, appearing on a TV show etc). To do so, we applied IE techniques to obtain semantic information from text.

In particular, this work focuses on handling texts in Portuguese for which it is necessary to deal with some unique aspects while specific language tools for Portuguese should be used.

The next section includes some background on IE and ontology population. Section 3 is a description of the dictionary structure. Section 4 presents the approach to this problem and the methods applied. Section 5 presents implementation issues and Section 6 summarizes some results. Finally, Section 7 presents related works and concluding remarks.

2 BACKGROUND

2.1 Information Extraction

Information extraction has been applied in the last years in different contexts and using different techniques. IE methods can be categorized in learning and knowledge engineering approaches (Kaiser and Miksch, 2005). Learning methods require an annotated corpus, which is usually expensive to obtain. Knowledge engineering methods use rules that are hand crafted and improved on an iterative process or obtained using an unsupervised technique.

Information extraction can be used to obtain entities (ontology classes), attributes (ontology classes' slots), facts (relations between classes) and events (activities or events where entities take part) (Feldman and Sanger, 2007).

Several tasks are necessary to accomplish IE, such as named entity recognition, anaphora resolution, coreference resolution, template element task, template relationship task, and scenario template production. Additionally, general-purpose natural language pre-processing steps are necessary, like tokenization, paragraph and sentence splitting, PoS (part-of-speech) tagging, shallow or full parsing etc.

2.2 Ontology Population

Ontologies have gained popularity in recent years due to the emergence of the semantic web, where ontologies play a central role. One of the bottlenecks of the semantic web is acquiring semantic content, i.e, semantically annotating the existing content on the web. Text content available on the web is primarily intended for human consumption which means that the semantics contained in texts is not explicitly represented but is in the reader's mind. Automatically (or semi-automatically) acquiring domain knowledge through ontology population is an important contribution towards semantic web applications. After extracting semantic information from text and representing it as a domain ontology, new applications arise, such as integrating information, navigating the knowledge base on a structured manner and finding new relationships not explicit from the source texts.

Ontology learning is the task that aims at discovering concepts and relations from text. In contrast, ontology population (OP) from text is the task of acquiring instances of concepts and their relations from text (Tanev and Magnini, 2006).

IE and OP are closely related in the sense that an ontology can be used to represent the IE process output and, on the other hand, knowledge represented in the ontology can help the IE cycle. This interaction between IE and OP is referred to as ontology driven information extraction (Yildiz and Miksch, 2007).

3 DICTIONARY DESCRIPTION

The Cravo Albin's Dictionary of Brazilian Popular Music (*Dicionário Cravo Albin da Música Popular Brasileira*) (Albin, 2008) is maintained by the Cravo Albin Cultural Institute, a civil, non-profit organization headquartered in Rio de Janeiro City and established in 2001. The dictionary contains more than 5.000 entries with information on biography, artistic data, shows, video clips and related bibliography of Brazilian popular music artists. It is available on the Internet, and it is organized and accessed by artist noun entries. It is possible to search for entries, but there is no way to

navigate from entry to entry, since there are no hyperlinks. Although the whole content of the dictionary is available on the Internet, it is organized much like as a paper dictionary, i.e., there is linkage between topics or browsing facilities through its contents. The only function available is a search by main entry (artist noun). Each entry is presented on one or more HTML pages.
Figure 1 shows the home page of the dictionary.

Figure 1: Dictionary's home page.

Information is split into the following sections: full name, birth date and place, death date and place, biography, artistic data, work, discography, shows, video clips, historical and artistic data, reviews and bibliography.
Figure 2 shows a biography detail page example.

Figure 2: Biography detail page example.

According to Abiteboul et al (2000) data can be classified, in terms of its structure, in three categories. Data can be structured (both schema and data types are known), semi-structured (schema or data type is known) and unstructured (schema and data type undefined). Typically, a web page contains data in the latter two categories.

3.1 Semi-structured Information

Some data as name, full name, birth place and year are presented as semi-structured information (although with variable completeness). Some examples are shown in figure 3.

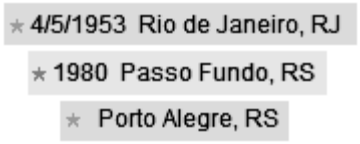

Figure 3: Semi-structured data represented as text.

The star icon indicates birth date and place. Note that completeness and precision may vary. The first example has a full birth date (4/5/1953), birth city (Rio de Janeiro) and birth state (RJ), while the second shows only the year of birth and the last has only city and state but no date information.

Information on discography, video clips, main shows and bibliography are presented on a semi-structured manner, as text, but following a standard format (structure and style). For instance, for discography section (*discografia*, in Portuguese), records are listed with publication year, record company and media type, like depicted in Figure 4.

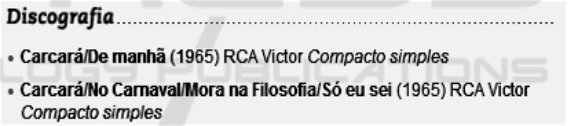

Figure 4: Semi-structured data represented as text.

3.2 Unstructured Information

Most of information is presented as free, natural language text, like in the artistic data and biography sections, as shown in Figure 5.

Maria Bethânia
Singer. Considered to be one of the greatest interpreters in the history of Brazilian popular music. Sister of the singer and composer Caetano Veloso. She began her artistic career in 1963, appearing in the play "Boca de Ouro" by Nelson Rodrigues.
...

Figure 5: Unstructured data, plain text.

The corpus used in this work is written in Portuguese, but some translated English examples are presented for the sake of readability.

4 ONTOLOGY POPULATION

Since the source information is a web site having no previous annotation, we have implemented an unsupervised semi-automatic method to generate extraction rules and templates using heuristics based on linguistic features, such as words and PoS tags.

4.1 Pre-processing

The first step was to fetch web pages and to identify HTML tags to separate sections from the dictionary.

During analysis of the page structure all necessary data to next steps were identified. Sections were identified through formatting marks, such as bullets and images. There was a one-to-one relationship between these marks used in HTML page and contents of the sections. These HTML tags were used together with regular expressions to determine sections boundaries. Table 1 shows some mapping examples.

Table 1: Mapping examples.

Pattern used on source	Mapped Content
HTML TAG <IMG SRC="img/bul_nasc.gif"> plus Regular Expression	Birth date Birth place
HTML TAG <IMG SRC="img/bio.gif"> plus Other section beginning or page end	Biography free text

Each dictionary entry is composed by one or more HTML pages that have a naming convention that was used to extract the corresponding content.

During the extraction from HTML, non-informative tags were filtered out in order to end with plain text corresponding to each section. Plain text was then loaded on a database to allow further processing.

During this stage special characters were also transformed to make them compliant to the linguistic tools that were used. For example, special UTF-8 open (“) and close (”) quotation marks were converted to ordinary quotation marks (").

4.2 Information Extraction

Information extraction requires some general purpose NLP processing like tokenization and domain specific tasks such as entity extraction. The following subsections describe the steps performed in the proposed method.

4.2.1 General Purpose NLP

Sentence splitting, tokenization and part of speech tagging were executed.

Named entities were classified (according to the domain ontology) in Artist (subclass of Person), Company (companies are divided into Record labels, Publisher companies and others) and Work (CDs, DVDs, Vinyl etc).

4.2.2 Temporal Expression Identification

Two types of temporal expressions were handled.

Type 1 is formed by expressions containing dates in numerical form, like “01/04/1970” or “1970, January”. This kind of expression always includes at least one numeric part. Nevertheless, many levels of precision and uncertainty may be expressed in this kind of expression, like “in the beginning of year 2000”, “circa 1870” or “in the middle of 1999”. These levels of precision and uncertainty were taken into account and represented on the ontology. Using the numeric expression as a trigger, and morphosyntactic rules, we constructed a list of patterns and classified each expression according to the ontology classes. These rules are based on word and PoS features. Table 2 shows some temporal expression patterns generated, indicating the chosen class and precision.

Table 2: Temporal expression pattern examples.

Pattern	Class	Precision
em <ANO> in <YEAR>	Instant	Year
entre <ANO1> e <ANO2> between <YEAR1> and <YEAR2>	Defined Interval	Year
a partir de <MÊS> de <ANO> from <MONTH> <YEAR> on	Interval, left open	Month
circa <ANO> circa <YEAR>	Around Instant	Year

Type 2 contains anaphors of type “in the following year” or “two years later”. This kind of expression must be resolved, linking them to the referent expression, in order to obtain a value for the temporal attribute. This is performed using a simple heuristic inspired in Mitkov’s algorithm (Chaves and

Rino, 2008). The nearest previous date expression identified as type 1 as described above was taken as the antecedent.

4.2.3 Artistic Works Identification and Classification

The main feature used for identifying artistic works (e.g. songs, shows, CDs etc.) was a verb related to an artistic activity, like the verbs to perform or to record, that are related to artistic activities. In particular, due to the dictionary's descriptive nature, verbs appear mainly in the past tense, such as *gravou* (recorded) or *lançou* (released). These verbs are used as triggers for frames defined to extract information on the object of the action (the patient, like a song or CD), who performed it (the agent, e.g., the artist involved in the action) and when such activity represented by the verb was performed (temporal adjunct). This task can be seen as a simplified approach to the semantic role labelling problem.

Figure 6 shows a frame example used to extract information for the verb *lançou* (released).

Verb (frame trigger)
lançou (launched)

Performer (agent)
In general, the dictionary entry is the performer, since most of the time, narrative is elliptical. Exception occurs when a named entity classified as an artist immediately precedes the verb *lançou*.

Temporal adjunct
Determined as described in section 4.2.2

Work (patient)
Taken from the named entity recognition step. Sentence structure analysis is performed to catch multiple works enumerated with commas and conjunctions.

Figure 6: Extraction frame sample.

In order to classify the artistic works, a gazetteer of work types and song genres was used in conjunction with morphosyntactic patterns. This gazetteer was constructed and pre-loaded on the ontology to support classification during information extraction.

4.2.4 Genealogical Relations

Genealogical relations were also managed using frames and a lexical database of expressions to detect genealogical relations, such as *filho de* (son of), *mãe de* (mother of) or *neta de* (granddaughter of). These linguistic patterns are linked to the corresponding ontology classes.

4.2.5 Partnership Relations

Partnerships (such as composing a song together or participating on the same record) were approached in a similar manner to genealogical relations.

4.3 Ontology Preparation

Music Ontology (Giasson and Raimond, 2007) was selected as a base ontology to represent the specific concepts of the music domain.

However, since some concepts necessary to represent the extracted information were not available on Music Ontology, a new, extended ontology was constructed based on it, adding new concepts and relationships.

From the Music Ontology the concepts of MusicArtist, MusicGroup, SoloMusicArtist, MusicalManifestation were used and translated to Portuguese to support linkage to the text content.

OWL-DL was chosen to represent the ontology, due to its sufficient expression power and tools available to deal with it (Cardoso, 2007).

In order to get insight on the most relevant terms (nouns and verbs) in this corpus, statistics on tokens tagged as verb or nouns were performed.

Table 3 shows the top 5 verbs and top 5 nouns (common names) found in the dictionary.

Table 3: Top 5 verbs and top 5 nouns.

PoS TAG	Word
V (Verb)	gravou (recorded)
	lançou (launched)
	participou (participated)
	apresentou (presented)
	interpretou (interpreted)
CN (Common Name)	ano (year)
	disco (record)
	samba
	música (song)
	parceria (partnership)

Due to the characteristics of events described in this dictionary, a time ontology is essential for capturing the semantic expressed in texts. Time is represented based on Allen's theory of time (Allen, 1991), where two kinds of entities exist: instant and interval. An instant is characterized by a "point" in time that can have different precision levels, like a

specific day, year or decade etc. An interval is defined by a starting instant and an ending instant. Furthermore, textual description of historical events usually contains some vagueness like "in the beginning of year 2000" or "circa 1890". Thus, the ontology must also be able to capture this kind of statements with vague concepts. (Mani and Wilson, 2000).

Ontology population from text demands handling imperfect information. Imperfection of information can be categorized as imprecision (vagueness), inconsistency (contradictory information) or uncertainty (trust level) (Haase and Völker, 2005).

In the dictionary's case, we dealt with imprecision, as we took the authority of the dictionary as trustable and assumed the content is consistent, since it is maintained by a dedicated group of music researchers.

4.4 Ontology Instantiation

After all the annotations described in section 4.2 were performed, the intermediate database was used to create instances of classes and relations described in the ontology. Before committing the extracted information to the ontology, a manual validation step was performed to ensure ontology consistency.

5 IMPLEMENTATION

To implement the described approach, a framework was developed based on a subset of the conceptual model proposed by Graça et al (2006). This framework enables processing text and having multiple annotating levels. It holds all the annotation performed during 1the IE process that is used afterwards to instantiate the ontology concepts and relations. An overview of the framework is depicted in figure 7.

Pre-processing task fetches and processes HTML pages and then store the output to the MySQL database. Information extraction tasks annotate text on multiple levels, like PoS tags, lemmas and named entities boundaries. During this phase, some information is also retrieved from the ontology, like gazetteers for work types and song genres, which were pre-loaded on the ontology by a separate process. The Ontology Instantiation step read annotated data from the database and creates classes and relations instances.

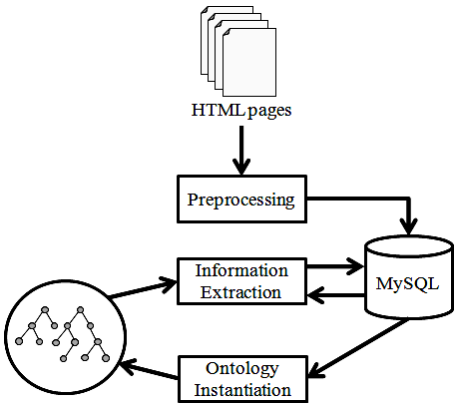

Figure 7: Framework overview.

The solution was implemented using MySQL 5.1 and programs were developed in Java and Python.

Paragraph splitting was performed based on formatting characteristics of the HTML pages. To perform sentence splitting, tokenization and PoS tagging, LX-Suite was used (Branco and Silva, 2006). LX-Suite is a freely available online service for the shallow processing of Portuguese.

To identify named entities, both structural hints (like parenthesis and quotation marks) and PoS tags output by LX-Suite (namely the PNM PoS tag for part of a proper name) were used.

Lemmatization was performed using the NILC lexical database (Muniz et al, 2005).

Protégé 3.3.1 (Protégé, 2008) was used to create the ontology. To instantiate concept and classes, the open-source Java library Protégé-OWL API was used (Knublauch, 2006).

6 RESULTS

In order to evaluate the information extraction process performance, a subset of 374 entries of the dictionary was used. This subset contains 9.912 sentences and 295,977 tokens.

From these sentences, 1,102 contain the verb *gravou* (recorded), 943 of them contain temporal adjuncts. The other ones have no time information, as they simple state facts without temporal properties, like "Cláudia recorded a vinyl record in Japanese".

Using the proposed heuristics to generate time expression extraction templates, a set of 33 different patterns was obtained.

Applying these patterns to this test set, 89.9% of time expressions were extracted (recall) and 86.3% (precision) were correctly mapped to the

ontology. This result corresponds to a F1 score of 88.1% for time expressions extraction.

Precision and recall was evaluated by manually tagging and inspection over this sample.

From a set of 260 sentences containing relative temporal expressions (anaphoric), 80.4% were correctly resolved using the proposed heuristic.

A small fragment of the populated domain ontology represented as a graph is shown in Figure 8. It shows classes for Work (Record) and Artist (SoloMusicArtist) and some of their corresponding instances.

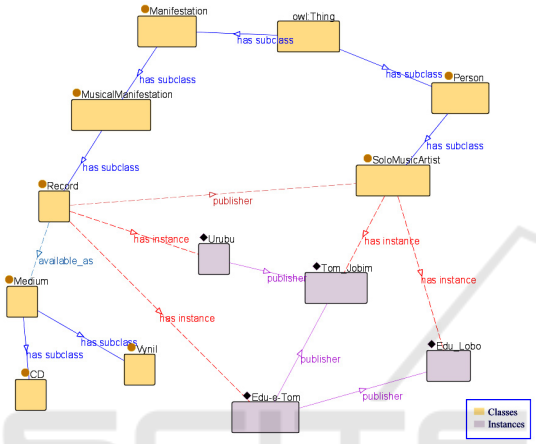

Figure 8: Ontology fragment sample.

7 DISCUSSION

This paper presented an unsupervised rule-based method for information extraction from Brazilian Portuguese texts that reaches a reasonable extraction performance with no need for training data. Although the herein proposed method is dependent on the domain and the source, it allows extracting relevant information to the domain without having to fully annotate the corpus, what would be required if a supervised method had been used. Other contributions of this work are dealing with Portuguese texts, and building a domain ontology for the popular Brazilian music history.

7.1 Related Work

Other unsupervised approaches for OP are described in (Hearst, 1992) and in (Cimiano and Völker, 2005). Hearst patterns were adapted in our method to consider also structural patterns and verb valence. Cimiano and Völker use a dependency parser which is not available for Portuguese.

A weakly supervised method is described by Tanev and Magnini (2006). Although it has better performance than the unsupervised approaches, it requires a training set.

7.2 Future Work

A potential use of the populated ontology is to support a semantic browser, where the original text from the dictionary can be visualized together with the semantic labels, like described in (Quan and Karger, 2004).

The populated ontology also permits visualize links not explicitly expressed in the texts, like people born at the same city or year. Navigating through these discovered dimensions can reveal interesting associations among artists.

Additionally, anchoring to other sources can expand information on the entries. For example, the CliqueMusic site (CliqueMusic, 2008) also contains information on Brazilian musicians, so that the ontology can be enriched (or at least linked to) the corresponding entries on the other database.

To deal with multiple sources, the ontology must be able to handle conflicting and different trust levels. So another path for evolution is to extend the ontology to support imperfect information, as discussed in (Haase and Völker, 2005).

Inference engines like the one described in (Haarslev and Möller, 2003) can be applied to the OWL ontology in order to derive new knowledge. Some examples include inference using genealogical relations, temporal overlapping (that can be used to determine that two artists were contemporaneous), geographical hierarchies like city contained in a state and so on.

Although in the case of dictionary it was not necessary due to relatively simple and regular HTML structure, another possible evolution is to implement a wrapper induction process to generalize HTML page extraction (Chang et al, 2006).

REFERENCES

Abiteboul, S., Buneman, P., Suciu, D., 2000. *Data on the Web*. San Francisco: Morgan Kaufman.

Albin, R., 2008. *Dicionário Cravo Albin da Música Popular Brasileira*, http://www.dicionariompb.com.br, accessed on November, 2008.

Allen, J., 1991. *Time and Time Again - The Many Ways to Represent Time*, International Journal of Intelligent Systems, 6 (1991).

Branco, A., Silva, J., 2006. *A Suite of Shallow Processing Tools for Portuguese:LX-Suite*, In *Proceedings of 11th*

Conference of the European Chapter of Association for Computational Linguistics, pp. 179-182.

Cardoso, J., 2007. *The Semantic Web Vision: Where are We*, IEEE Intelligent Systems, September/October 2007, pp.22-26, 2007.

Chang, C., Kayed, M., Girgis, M., Shaalan, K., 2006. *A Survey of Web Information Extraction Systems*, IEEE Transaction on Knowledge and Data Engineering, 18(10), pp.1411-1428.

Chaves, A. and Rino, L., 2008, *The Mitkov Algorithm for Anaphora Resolution in Portuguese*. In *International Conference on Computational Processing of Portuguese Language (PROPOR 2008)*, Aveiro, Portugal.

Cimiano, P. and Völker, J., 2005. *Towards large-scale open-domain and ontology-based named entity classification,* In *Proceedings of RANLP'05*, pp. 166–172, Borovets, Bulgaria.

CliqueMusic, 2008. *CliqueMusic site*, http:// cliquemusic.uol.com.br, accessed on November, 2008.

Feldman, R. and Sanger, J., 2007. *The Text Mining Handbook - Advanced Approaches in Analyzing Unstructured Data*, Cambridge University Press, Cambridge, UK.

Giasson, F. and Raimond, Y., 2007. *Music Ontology Specification*, http://musicontology.com, accessed on November, 2008.

Graça, J., Mamede, N., Pereira, J., 2006. *A framework for Integrating Natural Language Tools*, In *Computational Processing of the Portuguese Language – 7th International Workshop, PROPOR 2006*, Itatiaia, Brazil, Springer

Gruber, T., 2008. *Ontology*. To appear in *Encyclopedia of Database Systems*, Ling Liu and M. Tamer Özsu (Eds.), Springer-Verlag

Haarslev, V. and Möller, R., 2003. *Racer: An OWL Reasoning Agent for the Semantic Web*, In *Proceedings of the International Workshop on Applications, Products and Services of Web-based Support Systems, in conjunction with 2003 IEEE/WIC International Conference on Web Intelligence,* Halifax Canada, Oct 13, pp. 91-95, 2003.

Haase, P. and Völker, J., 2005. *Ontology learning and reasoning - dealing with uncertainty and inconsistency* In *Proceedings of the Workshop on Uncertainty Reasoning for the Semantic Web (URSW 2005)*

Hearst, M., 1992. *Automatic acquisition of hyponyms from large text corpora*, In *Proceedings of the 14th International Conference on Computational Linguistics (COLING)*, pp. 539-545.

Kaiser K., and Miksch, S., 2005. *Information Extraction – A Survey*, Technical Report Asgaard-TR-2005-6, Vienna University of Technology, Vienna, Austria, 2005

Knublauch, H. 2006. *Protégé-OWL API Programmer's Guide*, http://protege.stanford.edu/plugins/owl/api/ guide.html, accessed on November, 2008.

Mani, I., and Wilson, G., 2000. *Temporal Granularity and Temporal Tagging of Text*. In *AAAI-2000 Workshop on Spatial and Temporal Granularity*, Austin, TX.

Moens, M-F., 2006. *Information Extraction: Algorithms and Prospects in a Retrieval Context*, Springer.

Muniz, M. and Nunes, M., Laporte, E., 2005. *UNITEX-PB, a set of flexible language resources for Brazilian Portuguese* In *Proceedings of the Workshop on Technology on Information and Human Language (TIL)*, São Leopoldo, Brazil

Protégé, 2008. Protégé home page, http:// protege.stanford.edu/, acessed on November, 2008.

Quan, D. and Karger,D., 2004. *How to make a semantic web browser,* In *Proceedings of the 13th international conference on World Wide Web,* 2004

Tanev, H. and Magnini, B., 2006. *Weakly Supervised Approaches for Ontology Population* In *Proceedings of 11 th Conference of the European Chapter of the Association for Computational Linguistics: EACL 2006*

Yildiz, B. and Miksch, S., 2007. *Motivating Ontology-Driven Information Extraction*, In *Proceedings of the International Conference on Semantic Web and Digital Libraries (ICSD-2007).*