FALA 2010
VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop

Knowledge Extraction from Minutes of Portuguese Municipalities Meetings

Mário Rodrigues[1], *Gonçalo Paiva Dias*[2], *António Teixeira*[3]

[1]ESTGA/IEETA, [2]ESTGA/GOVCOPP, [3]DETI/IEETA
University of Aveiro, Aveiro, Portugal
mjfr@ua.pt, gpd@ua.pt, ajst@ua.pt

Abstract

A very relevant problem in e-government is that a great amount of knowledge is in natural language unstructured documents. If that knowledge was stored using a computer-processable representation it would be more easily accessed. In this paper we present the architecture, modules and initial results of a prototype under development for extracting information from government documents. The prototype stores the information using a formal representation of the set of concepts and the relationships between those concepts - an ontology. The system was tested using minutes of Portuguese Municipal Boards meetings. Initial results are presented for an important and frequent topic of the minutes: the subsidies granted by municipalities.

Index Terms: entities and relations extraction, e-government, semantic query.

1. Introduction

E-government relates to the use of information and communication technologies (ICT) by government, including the online provision of government services. These technologies have the potential to improve the government service delivery, can empower citizens through access to information, or make government management more efficient. A great amount of information in local and central government agencies is registered in unstructured formats, making the access/query/search to it not readily available. Although many of these documents are stored in computers, their format prevents the information they contain to be computer-processable and thus they cannot be manipulated to meet user's specific needs. E-government would benefit from systems able to integrate several sources of information and able to understand unstructured documents [1].

The minutes of Municipal Board meetings contain information that would benefit from being made available in searchable knowledge bases. These documents are important because municipalities are often the closest point of service for citizens and enterprises, and the minutes record the decisions of the Municipal Board.

In this paper we present a system able to create semantic information from this type of documents - natural language, unstructured - using natural language processing algorithms, integrating open source software, and using external sources of information as Google Maps and Geo-Net-PT01. The remaining of the paper starts by discussing the related work. Section 2 starts with an overview of the developed system and continues by elaborating on each of the three parts that compose it. In Section 3 the initial results are presented and discussed. The paper ends with the conclusions in Section 4 and the acknowledgments in Section 5.

1.1. Related Work

The research activity done so far in e-government is usually centered in solving problems as interoperability and service integration, which are very important problems and should be further addressed. In such projects it is usually considered that the information is already in the system, whether placed by human operators or using existing databases (e.g. OneStopGov and Access-eGov). To our knowledge, no project was dedicated to the relevant problem of automatic acquisition of information from natural language government documents [1].

Several projects were dedicated to the task of scalable, domain independent information extraction (IE). Some are more focused in building semantic knowledge bases from Wikipedia - DBpedia, Kylin, YAGO/NAGA and others. They generally extract information from the (natural language) documents texts and use the Wikipedia's structure to infer the semantics. Other state of the art systems work at the web scale - KnowItAll, TextRunner. Finally, some works focus on the knowledge evolution over time - TARSQI, Timely YAGO.

2. Developed System

The proposed system reuses open source software - adapted to work with Portuguese - to take advantage of the state of the art approaches and software. Specific software was developed to integrate the reused software in a coherent system (Figure 1).

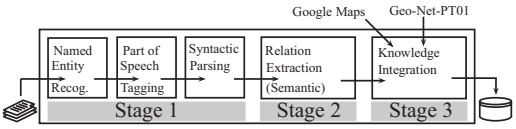

Figure 1: *The three stages of processing and their components. Input: natural language; Output: structured information.*

This information extraction (IE) system is composed of three parts, organized in pipeline architecture. First, the text is extracted and enriched with the inclusion of named entity (NE), part-of-speech (POS) tags and syntactic information. Second, the system uses this enriched text to train a classifier that looks for patterns of information in order to detect relationships between entities. Third, the system integrates the knowledge extracted in the previous part with information from other sources (e.g. geocode information) and stores it in a knowledge base that conforms to a defined ontology. This part also includes the capabilities to semantically query the knowledge base.

The following sections explain how each module was adapted and integrated. The explanations are illustrated using the fragment of a minute presented in the first row of Table 1.

Fragment:	"Seguidamente, a Câmara deliberou, por unanimidade, atribuir os seguintes apoios financeiros: ... À ARCEL - Associação Recreativa e Cultural de Espinhel, um subsídio no valor de 8.640,00€, destinado a apoiar a execução do Plano Anual e a Escola Artística"
Translation:	"Subsequently, the Board decided unanimously to award the following financial aid: ... To ARCEL - Associação Recreativa e Cultural de Espinhel, a subsidy amounting to €8,640.00, to support the implementation of the Annual Plan and the Art School".
NER out:	<EM C1="EM">À ARCEL</EM>- <EM C1="ORGANIZATION" C2="INSTITUTION">Associação Recreativa e Cultural de Espinhel</EM>, <EM C1="NUMBER" C2="TEXTUAL">um </EM>subsídio no valor de <EM C1="VALUE" C2="MONEY">8.640,00€</EM>, destinado a apoiar a execução do <EM C1="OBRA" C2="PLAN">Plano Anual</EM> a <EM C1="EM">Escola Artística</EM>.

Maltparser:

1	À_Arcel	À_Arcel	prop	prop	_	0	UTT	_ _
2	–	-	punc	punc	_	1	PUNC	_ _
3	Associação_Recreativa...	Associação...DE_Espinhel	prop	prop	_	1	N<PRED	_ _
4	,	,	punc	punc	_	1	PUNC	_ _
5	um	um	art	art	_	6	>N	_ _
6	subsídio	subsídio	n	n	_	1	N<PRED	_ _
7	em	em	prep	prep	_	6	N<	_ _
8	o	o	art	art	_	9	>N	_ _
9	valor	valor	n	n	_	1	N<PRED	_ _
10	de	de	prep	prep	_	9	N<	_ _
11	8.640,00€	8.640,00€	num	num	_	9	N<	_ _
	(...)							

LEILA out:	503.0689326929 a_arcel 8.64000eur # 20091112171306990926.leilaout, Bridge: #3: <first> -unknown-> @dummy@ -unknown-> <second> EXAMPLE
KB entry:	<owl:NamedIndividual rdf:about="http://mri.ieeta.pt/2010/04/07/municipality.rdf#s_8.64000eur"> <rdf:type rdf:resource="http://mri.ieeta.pt/2010/04/07/municipality.rdf#Subsidy"/> <moneyAmount>8.64000eur</moneyAmount> <assignedTo rdf:resource="http://mri.ieeta.pt/2010/04/07/municipality.rdf#a_arcel"/> <terms:isReferencedBy rdf:resource="http://mri.ieeta.pt/2010/04/07/municipality.rdf#acta_20091112171306990926"/> </owl:NamedIndividual>

Table 1: Example fragment and output of several system modules.

2.1. Text Annotation/Enrichment

This section describes the modules that compose the first part of the system: named entity recognition (NER), POS tagging, and syntactic parsing.

2.1.1. Named Entity Recognition

The current version of NER is based on a system developed for Portuguese named Rembrandt. It uses Wikipedia as a raw knowledge resource and its document structure to classify all kinds of NEs in the text [2]. An early version of this module, designed to detect locations, was based on a small set of rules and the geographical ontology Geo-Net-PT01 [3, 4].

Rembrandt tries to classify each NE according to the Second HAREM directives [5]. Unclassified NEs are collected to be classified using other strategies. For now, the strategy is to query the Google Maps API to have the location of the NE and, if a location is retrieved, to classify it as an entity with a fixed physical location. In this case, the entity is marked as having latitude and longitude which can be an organization (enterprise or institution headquarters), a place (physical or human), or an event that happens always in the same place.

Regarding our example, the output of Rembrandt can be found in the third row of Table 1. The example shows that Rembrandt has identified six NEs and was able to classify four of them. The class is the value of C1 in the tag EM. When that value is EM (C1="EM") it means that the NE was not classified.

2.1.2. Part-Of-Speech Tagging

The POS tagging is performed by TreeTagger. It annotates text with POS and lemma information and has been successfully used to tag several natural languages including Brazilian Portuguese. TreeTagger implements a decision tree to obtain reliable estimates of context transition probabilities in order to avoid sparse-data problems, a relevant problem in statistical models training. The decision tree automatically determines the appropriate size of the context - number of surrounding words - which is used to estimate the transition probabilities [6].

The tagger was trained with a European Portuguese lexicon in order to be integrated in the system. The tagger provides tools to train a language model given three files: a corpus with tagged training data; a full form lexicon; and an open class file with the list of possible tags of unknown word forms.

The corpus used to train it - and the syntactic parser - was Bosque v7.3, the only Portuguese corpus usable to train the chosen syntactic parser. Bosque is a subset of Floresta (a publicly available treebank for Portuguese), fully revised by a linguistic team, that contains about 185,000 words [7]. The full form lexicon used in the training process was based on the computational lexicon LABEL-LEX-sw that comprises more than 1,500,000 inflected word forms, automatically generated from a lexicon of about 120,000 lemmas [8]. The output of the tagger is in the fourth (and fifth) column of the fourth row of Table 1.

2.1.3. Syntactic Parsing

The syntactic parsing is done with a data-driven dependency parser named MaltParser [9]. MaltParser was selected because there are no parsers freely available for Portuguese. It can be used to induce a parsing model from treebank data - and to parse new data using that model - and was already successfully used to parse several natural languages as English, French, Greek, Swedish, and Turkish. The parsing algorithm used was the same of the Single Malt system [9].

The parsing model was induced with the version 7.3 of

Bosque used in the Tenth Conference on Computational Natural Language Learning (CoNLL-X) shared task: multi-lingual dependency parsing. This particular version was selected because, as far as we know, is the only version - and treebank - that is available for Portuguese in the CoNLL-X format, the format accepted by MaltParser. The format defines a sentence as one or more tokens, each one starting in a new line and consisting of ten fields.

The output of the parser is presented in the fourth row of the Table 1. The value of the seventh field - HEAD - is assigned by the parser and indicates the head of the current token. For instance, the HEAD field of the 11th token (8.640,00€) is 9, which means that "8.640,00€" depends syntactically of "valor" - the 9th word which means value.

After the parsing step, the POS tag assigned to each NE is replaced by the class assigned by Rembrandt to that NE. This allows taking advantage of the information given by Rembrandt about the class of the NEs.

2.2. Relation Extraction

The general problem of interpreting text involves the determination of the semantic relations among the entities and the events they participate in. Informally, this task aims to detect elements as "who" did "what" to "whom", "when" and "where" [10].

The approach followed was to use LEILA. LEILA is a system able to extract instances of arbitrary binary relations [11]. It was chosen because it uses deep syntactic analysis to detect the relations in natural language sentences. The syntactic analysis of the original configuration is performed by a link grammar parser. The linguistic structures constructed by this parser are connected planar undirected graphs called linkages. The words of the sentence are the nodes of the graph and the edges are called links and have labels. An adapter was build for LEILA to be able to use the output of MaltParser. The adaptation was mostly straightforward and a linkage of the example sentence is represented in Figure 2.

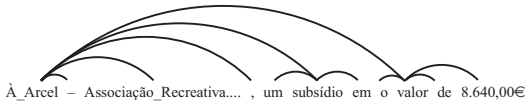

Figure 2: *The linkage of the syntactic structure of the fragment presented in Table 1.*

Each target relation of LEILA requires a function to decide into which category a pair of words (entities) falls. The pair can be: an *example* if it belongs to a list of examples; a *counterexample*, deducted from the examples if it is incompatible with the examples; a *candidate* if it obeys to some criteria and is neither an *example* nor a *counterexample*; or can be none of the previous and should be ignored. Using the output of the function and one classifier - it was used k-nearest neighbor - the core algorithm has three phases: in the *discovery phase* seeks linkages where the example pairs appear to produce positive patterns, and collects as negative patterns all linkages that match a positive pattern but produce a counterexample; in the *training phase* it uses statistical learning to produce a pattern classifier based on the patterns acquired in the discovery phase; in the *testing phase* the classifier evaluates all sentences: if a pair of entities is classified as *candidate* and the pattern connecting the entities is classified as positive, the pair is considered a new element of the target relation.

For the proof of concept and first application of the system a function was developed to detect subsidies granted by municipalities. This subject of was selected because the amount of subsidies (granted to whom) is a relevant issue in local government. The output of LEILA is at the fifth row of Table 1.

2.3. Information Integration, Management and Access

An ontology was created to define the semantics of the knowledge base. To be as standard as possible the ontology results from the usage of well known ontologies with a minimum amount of entities and properties added. It combines the ontologies Friend of a Friend (FOAF), Dublin Core, World Geodetic System (1984 revision), and GeoNames (full version), with a new class (Subsidy), a new object property (assignedTo), and a new data type property (moneyAmount).

A reasoner checks the coherence between new information and the information already in the knowledge base. The new information is added to the knowledge base when it is coherent with the existing one. Otherwise is discarded (for the moment). The reasoning is performed by an open source reasoner for OWL-DL named Pellet. It supports reasoning with individuals and user defined data types [12].

2.3.1. Geo Location

The information about entities with a fixed location (as streets, organizations headquarters, and some events) is enriched with its geocoding information. The geocodes are obtained via queries using the Google Maps API. The political organization of the spaces - street $\subset$ neighborhood $\subset$ city $\subset$ municipality ... - is obtained using a free geographic ontology of Portugal with about 418,000 features named Geo-Net-PT01 [3]. This allows the system to display the information spatially on a map and to search and relate information by its location.

2.3.2. Knowledge Base

The knowledge base is defined with the web ontology language (OWL). The storage and management is performed by Virtuoso Universal Server which features an endpoint for SPARQL.

The last row for Table 1 shows an entry of the knowledge base relative to the example. It is visible an *owl:NamedIndividual* of type *Subsidy* with some *moneyAmout*, with property *assignedTo a_arcel* and *isReferencedBy acta_2009...* (in Portuguese "acta" means minute). This entry defines a subsidy. Other existing entries (not showed here) define the minute *acta_2009...* and the NE *a_arcel*.

3. Initial Results

First Experiments were performed to extract information (for now just subsidies) from several documents belonging to a Portuguese municipality, Águeda town municipality. The experiments were conducted using 23 minutes of the Municipal Board meetings. Information regarding subsidies was extracted and subject to query and evaluation against manual annotations. Each subsidy annotation identifies the entity which received the subsidy and the amount of money involved.

3.1. Examples

It is possible to make complex queries to retrieve information not explicit (or not present) in the documents. For example, it is possible to query which Institution had the highest number of subsidies or the highest total value, or the distribution of

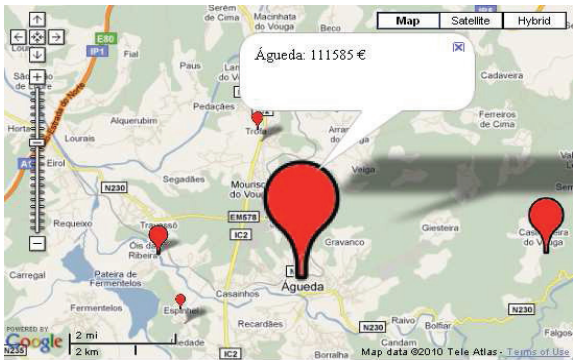

Figure 3: *Total amount of subsidies per "freguesia". The area occupied by the marker is proportional to the amount of money received. Five "freguesias" of Águeda municipality had subsidies for associations/companies/schools in the selected minutes: Águeda; Castanheira do Vouga; Espinhel; Óis da Ribeira; and Trofa.*

the total amount of subsidies granted by administrative subdivisions of the Municipality. The later was queried to the knowledge base. The query involves retrieving all subsidies of all entities which are located in one *freguesia* (neighborhood) of Águeda, group them by *freguesia* and compute the total amount per *freguesia*. After this query, the geographical location of the *freguesias* was added, and a web interface renders the result in a map as showed in Figure 3.

3.2. Performance Evaluation

A total of 107 subsidies were manually annotated in the test set. In the same set the system detected 62 subsidies - 53 well detected. The performance of the system was measured against the manual annotations. It was also measured discounting the subsidies found in enumerations (36) because enumerations like "the following subsidies were granted: *entity1 - amount1*; *entity2 - amount2*;..." were not considered for now. To detect what kind of thing is being enumerated it is necessary to have the context because enumerations (can) span across several sentences and the current algorithm processes one sentence at a time. It is being currently studied the way context tracking can be seamlessly integrated in the system. Table 2 summarizes the results.

Table 2: *The performance of the system measured for 23 documents containing 107 subsidies. Results were also compared against all subsidies that were not in enumerations (71).*

	detections				
	true	false	precision	recall	F_1
all (107)	53	9	0.85	0.50	0.63
not enum. (71)	53	9	0.85	0.75	0.80

The system achieved a precision (prec.) of 0.85 and a recall of 0.50 or 0.75 when enumerations are not considered. This performance is comparable to state of the art systems: DBpedia (prec. 0.86 to 0.99; recall 0.41 to 0.77), Kylin (prec. 0.74 to 0.97; recall 0.61 to 0.96), and YAGO/NAGA (prec. 0.91 to 0.99; recall not reported).

4. Conclusions

This article presented a first version of an IE system for natural language (government) documents in Portuguese. In this early stage of development the goal was to make a proof of concept and to perform a first evaluation. IE for e-government and in Portuguese is still a challenge because it requires research and adaptation to the specific area and language. Such systems are important because e-government own success can depend on how easy it is to maintain and use its services.

This design can be seen as a framework where different modules can be plugged in and/or switched. The current version is capable of acquiring information about subsidies and to integrate it with geographical information. The knowledge base complies with an ontology resulting form the merge of 4 public, mature, and broadly supported ontologies thus increasing interoperability. Initial results seem promising. They showed a system performance comparable with state of the art systems.

To conclude, the system works for Portuguese and was built reusing state of the art third party software, mostly developed aiming the English language. This shows that it's possible (and should be further tempted) to integrate high performance software tools designed for other natural languages.

5. Acknowledgements

The authors would like to thank Câmara Municipal de Águeda for making their documents available in digital format, and to thank Carlos Pereira for the careful annotation of them.

6. References

[1] M. Rodrigues, G. P. Dias, and A. Teixeira, "Human Language Technologies for E-Gov," in *Proc. of the 6th WEBIST*, 2010.

[2] N. Cardoso, "REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto," in *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM.* Linguateca, 2008.

[3] M. Chaves, M. Silva, and B. Martins, "A Geographic Knowledge Base for Semantic Web Applications," in *Proc. of SBBD*, 2005.

[4] M. Rodrigues, G. P. Dias, and A. Teixeira, "Automatic Extraction and Representation of Geographic Entities in eGovernment," in *Proc. of the 5th CISTI*, 2010.

[5] C. Mota and D. Santos, Eds., *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM.* Linguateca, 2008.

[6] H. Schmid, "Probabilistic Part-of-Speech Tagging Using Decision Trees," in *Proceedings of International Conference on New Methods in Language Processing*, vol. 12. Manchester, UK, 1994.

[7] C. Freitas, P. Rocha, and E. Bick, "Floresta Sintá (c) tica: Bigger, Thicker and Easier," *Computational Processing of the Portuguese Language*, 2008.

[8] E. Ranchhod, C. Mota, and J. Baptista, "A Computational Lexicon of Portuguese for Automatic Text Parsing," in *Proc. of SIGLEX99: Standardizing Lexical Resources - ACL*, 1999.

[9] J. Hall, J. Nilsson, J. Nivre, G. Eryiğit, B. Megyesi, M. Nilsson, and M. Saers, "Single Malt or Blended? A Study in Multilingual Parser Optimization," in *Proc. of the EMNLP-CoNLL*, 2007.

[10] L. Màrquez, X. Carreras, K. C. Litkowski, and S. Stevenson, "Semantic Role Labeling: An Introduction to the Special Issue," *Computational Linguistics*, vol. 34, no. 2, 2008.

[11] F. Suchanek, G. Ifrim, and G. Weikum, "LEILA: Learning to Extract Information by Linguistic Analysis," in *Proc. of the ACL Workshop OLP*, 2006.

[12] E. Sirin, B. Parsia, B. C. Grau, A. Kalyanpur, and Y. Katz, "Pellet: A Practical OWL-DL Reasoner," *Web Semantics: science, services and agents on the World Wide Web*, vol. 5, no. 2, 2007.