Economic Activity Recognition

António Paulo Santos[1], Paulo Bengala[2], and Nuno C. Marques[3]

[1] GECAD, Institute of Engineering - Polytechnic of Porto, Portugal
[2] Inspirennovit - Lda, Portugal
[3] CITI-DI-FCT, Universidade Nova de Lisboa, Monte da Caparica, Portugal
pgsa@isep.ipp.pt, paulo.bengala@inspirennovit.com, nmm@fct.unl.pt

Abstract. The identification of the economic activities performed by a company and its recognition from the text in the company's web site, is a task that has not yet received much attention in text mining and business intelligence applications. In this paper, we present a system designed for recognising economic activities performed by companies from text obtained from their websites. The applied strategy makes use of the composition of an economic activity, which reduces the size of the required gazetteers. The system achieved a recall of 63%, a precision of 61%, and an F1-measure of 62%.

Keywords: activity recognition, information extraction, business intelligence

1 Introduction

Knowing the economic activities performed by a company may have important applications in the Business Intelligence field. Basically, may help to characterize a company. This may have different applications, such as in joint ventures [1], in the internationalization of companies [11]. The work described in this paper is a module of a larger system having another application: the clustering and location of potential suppliers.

A way of determining the economic activities performed by a company is to extract this information from its website. This raises the problem of the *economic activity recognition*. The task is illustrated in the following examples using text from different company websites:

1. **"O Grupo Portucel Soporcel é hoje líder na Europa na [$_{ACTIVITY}$ produção de papéis de escritório]..."**
   (*The Portucel Soporcel Group is today Europe's leading in the [$_{ACTIVITY}$ production of office paper]...*)
2. **"A 2AB iniciou a sua actividade em 1987 tendo como principal actividade o [$_{ACTIVITY}$ fornecimento de equipamentos para a construção civil]..."**
   (*The 2AB started its activity in 1987 having as main activity the [$_{ACTIVITY}$ supplying of equipment for the construction industry]...*)

The economic activity recognition problem (defined on section 2.1) is challenging because there are many ways of expressing an economic activity and because of the lack of annotated data. Decomposing an economic activity into semantic classes allowed us to alleviate with those variations (section 2.3). The rule based technique plus the gazetteers helped to mitigate the lack of annotated data.

This paper presents the development of a real business system for recognising economic activities from text in Portuguese (sections 2.2, 2.3, 2.4, 2.5). The recognised activities can then be extracted and inserted into a database. This paper contributes to the Knowledge Discovery field applied to Business Intelligence by describing the experience and lessons learned during the development of a practical economic activity recognition system (section 3). A key element of this system is a set of text extraction rules that identify relevant information to be extracted.

2 Economic Activity Recognition

2.1 The Task

For the purpose of this work, we define an economic activity as any activity that can be performed in exchange for money or money's worth. Some examples include: *production*, *distribution* and *sale* of *any goods and services*.

We define the economic activity recognition task as: given a sentence, detect where an economic activity begins and ends. Example:

1. **"...é uma empresa dedica à [$_{ACTIVITY}$ produção e transporte de energia elétrica]."**
   (*"...is a company dedicated to the [$_{ACTIVITY}$ production and transport of electric energy]."*)

Note that recognised activities cannot overlap. Therefore, strings like *[production and transport of electric energy]* are identified as a single ACTIVITY, even representing two activities. The *"production of electric energy"* and *"transport of electric energy"* activities.

Finally, we define our task as: given a company's website url, identify the *economic activities performed* by it. Therefore, our task is to **recognise** the economic activities and **classify** them as *performed* or *not performed*.

2.2 The Approach - Overview

Given the company's website URL, we start by making use of what we call **external clues**, to select which web pages to visit and extract text. They are called *external clues* because they are not into the text that will be used by the system during the *activity recognition* and *classification* phases. The main idea behind an external evidence is as follows. Based on authors' experience, we know that not all web pages of a company website carry information about its

economic activities. For example, a "*contact*" or a "*site map*" page, usually do not contain information about the company's activities. On the other hand, it is very frequent to find the company's activities on a page, such as the "*about us*" page. Therefore, after entering in the company's home page, we try to locate this kind of pages. This is done by seeking for links (*HTML <a> tag*) containing text such as "*about us*" as done by Horacio Saggion et al [11]. Moreover, we also seek on the "*href*" attribute of <a> tag (e.g. *<a href="about.php">*), and on links containing images. On images we seek on the *src* and *alt* attributes (e.g. *<img src="about.gif" alt="about the company">*).

After extracting the text from the potential relevant web pages, the task of recognising *economic activities performed* by a company is decomposed into two phases:

1. **Economic activity boundary recognition.**
   The goal of this phase is recognising where an economic activity begins and ends in a sentence. For example, finding: *"conservação de arte moderna"* (*modern art conservation*) and not just *"conservação de arte"* (*art conservation*).
2. **Classification of the economic activity as "performed" or "not performed".**
   The goal of this phase is distinguishing between activities that are performed by the company and activities mentioned in a different context.

We start by describing the main ideas behind the strategy for recognising *economic activities* (section 2.3) and for classifying them as *"performed"* or *"not performed"* (section 2.4), describing finally the system architecture (section 2.5).

2.3 Phase 1 - Economic Activity Recognition

A simple approach for recognising activities would be to use a list of activities (a gazetteer). This list could be built, for instance, from a source such as CAE Rev3[4] and then expanded. This approach would be simple, fast and independent of language in the sense that a system that would implement this approach would not need to change its architecture. However, would have disadvantages such as the development and maintenance of the activities list, could not deal with economic activities variants, and could not resolve ambiguity.

Our approach makes use of the **composition of an economic activity** and **contextual clues** found in a sentence, overcoming the above disadvantages. The **composition of an economic activity** is based on the principle of the compositionality (sometimes called *Frege's Principle*), which states that "the meaning of a complex expression is derived from its parts" [10]. The fundamental philosophy behind *composition of an economic activity* is as follows. Economic activities such as *"fornecimento de equipamento para a construção civil"* (supply

[4] CAE Rev3 (Classificação Portuguesa de Actividades Económicas, Revisão 3) http://www.ine.pt/xportal/xmain?xpgid=caerev3&xpid=INE

of equipment for construction industry) are composed by a set of strings which suggests that they are economic activities. The previous example can be seen as composed by the string *"fornecimento"* (supply) which we will call an **ACTION**, the string *"equipamento"* (equipment) that we will call a **PRODUCT**, and *"construção civil"* (construction industry) an **ECONOMIC SECTOR**. Therefore an economic activity can be seen as a sequence of strings belonging to certain semantic classes of words. Those classes then may or may not be present in an economic activity, can occur in several orders, and can occur on adjacent and nonadjacent positions. The semantic classes identified in this work which can be part of an economic activity, are shown on Table 1. As shown on Table 1, some strings could belong to more than one semantic class (e.g. *energy* could be a SECTOR but also a PRODUCT).

Table 1. Semantic Classes of an economic activity

Semantic Class	**Meaning**	**Examples**
ACTION	Something that can be done or performed by a company. They are usually verbs or names derived from verbs.	produção (production), desenvolvimento (development), distribuição (distribution), armazenamento (storage), venda (sale).
PRODUCT	Products, services, product and service categories.	edifícios (buildings), energia (energy), arte (art), equipamentos (equipment).
SECTOR	Economic activities sectors and nom-compositional economic activities.	agricultura (agriculture), energia (energy), telecomunicações (telecommunications).

The approach makes also use of **contextual clues**. The idea is similar to what McDonald[7] called external (contextual) evidence, for identifying and classifying proper names. The fundamental philosophy behind a contextual clue within a sentence is as follows. Within a sentence, there may be a string which reinforces the presence of one or more economic activities or sectors in it. Currently we are using a particular type of clue. We are using keywords which may indicate that the following strings may be an economic activity or sector (e.g.: *activity*, *domain*, *sector*, *market*, etc.). We call these keywords SECTOR KEYs.

2.4 Phase 2 - Economic Activity Classification

After recognising the potential economic activities, the system should identify those that are performed by the company ignoring the remaining activities. For that we applied an approach based on the co-occurrence of certain strings and economic activities within the same sentence. The main idea is that the presence of certain strings in the same sentence as an economic activity may suggest that: 1) the activity is performed; 2) the activity is not performed or even isn't an economic activity in the particular context of the sentence. For example, if the

string *"policy"* co-occurs in the same sentence as *"data protection"*, from a text extracted from a company website, we will probably find the *"data protection policy"* of the company instead of the activity of *"data protection"*.

For applying the above idea, the system should use a "white" and a "black" list of strings. When analysing the co-occurrence of a potential economic activity and one or more of those strings in the same sentence, there are the following scenarios:

1. Only words of the "white list" co-occur with the economic activity. In this case, the activity should be classified as *"performed"*;
2. Words of both lists co-occur with the economic activity. In this case, the system should measure two distances in terms of the number of words. It should measure the distance between the activity and the closest word of the white list, and do the same for the closest word of the blacklist. The activity should then be classified as *"performed"* only if the closest string is in the white list;
3. In any other case the activity should not be classified as performed and therefore ignored.

2.5 The system Architecture

This section presents the system, based on the ideas presented in the previous sections. The system first gathers the main page of a company website and from there follows and gathers all potentially relevant pages. From the resulting web pages and using the GATE[5] [4] (*General Architecture for Text Engineering*) platform, the text is extracted from the web page (HTML document), then the text is splitted into tokens (words, numbers, etc.) and into sentences. Finally, taking as input the sentences and a set of lists, the system returns a set of *performed economic activities* for the company in question, based on the following steps:

STEP 1. Identification of the semantic classes
In this step, the system uses the collection of gazetteers for labelling words, one for each semantic class and one for SECTOR KEYs. The current version uses:

- A list of 22 SECTOR KEYs. A list of 105 ACTIONs;
- A list of around 23600 PRODUCTs (and services), from those 14200 are products and services from the *Nice classification*[6], and the remaining from the CPV (Common Procurement Vocabulary);
  Note that, although this is not true, we are assuming that every CPV entry is a PRODUCT. For example, the CPV entry *"45212212-5 Construção de piscinas"* (*Construction work for swimming pool*) it is an economic activity.
- A list of around 1200 SECTORs (and activities) from CAE Rev 3.

[5] http://gate.ac.uk/
[6] Classificação de Nice – Lista de Produtos e serviços (10ª Edição) http://www.marcasepatentes.pt/

It is important to note the small size of the lists used on the current system. In future versions, it would be desired to expand the above lists. Even the biggest list, the list of PRODUCTs with around 23600 entries is still small, given the wide variety of products and services that can be found in an economic activity.

STEP 2. Identification of Economic Activities
In this step, the system applies a set of syntactic rules based on the strings tagged as ACTION, PRODUCT, SECTOR, and SECTOR KEY on step 1, for recognising economic activities.

Table 2 shows the main rules for recognising a single economic activity. For example, rule 1 will recognise an activity composed by a single ACTION that may or may not be followed immediately by the word "*de*" (*of*) and ending in a PRODUCT or SECTOR (e.g. [$_{ACTION}$ *production*] of [$_{PRODUCT}$ *biodiesel*], [$_{ACTION}$ *rental*] of [$_{PRODUCT}$ *construction equipment*]).

Table 2. Rules for recognising a single economic activity composed by adjacent semantic classes. ("*de*" ("*of*") represents that word, <*word*> represents a single word, "*?*" means that the preceded symbol is optional, {*0,2*} means 0, 1, or 2 occurrences of the preceded symbol, and the vertical bar "|" is used for indicating a choice).

RID	Rule
1	ACTION "de"? (PRODUCT|SECTOR)
2	ACTION "de"? PRODUCT <word>{0,2} SECTOR
3	ACTION <word>? SECTOR_KEY <word>? (SECTOR|PRODUCT)

Table 2 illustrate the main rules for recognising a single economic activity composed by sequences of adjacent semantic classes. For each one of those rules, there is a version for dealing with non-adjacent semantic classes. The difference between the rules shown on Table 2 and those rules is that, *"de"?*, <*word*>*?*, and <*word*>{*0,2*} are all replaced by <*word*>*?*{*0,5*}. For example, rule 1 will not recognise the following string as an economic activity, but its variant will: "[$_{ACTION}$ *commercialization*] *of a wide range of* [$_{PRODUCT}$ *toys*]".

There is another set of rules similar to those presented Table 2, but for dealing with sequences of more than one occurrences of the same semantic class. For example, for recognising activities such as "[$_{ACTION}$ *collection*], [$_{ACTION}$ *storage*], and [$_{ACTION}$ *processing*] of [$_{PRODUCT}$ *blood*]".

Finally, there are a few syntactic rules based on SECTOR KEYs. For example, the rule "*SECTOR_KEY d[eoa]s? (PRODUCT|SECTOR)*" (where *d[eoa]s?* means "*of*"), which allows to recognise economic sectors or economic activities, such as "[$_{SECTOR\_KEY}$ *sector*] of [$_{SECTOR}$ *health*]". If there are multiple possible rules that match a string, the longest possible match is chosen.

At the end of this step the system has a set of potential economic activities. However, since its goal is to find economic activities performed by the company, two more steps are performed for classifying the activities as "*performed*" or

"*not performed*".

Step 3. Identification of "performed" terms
In the current implementation, the system uses a white list (manually created) of about 60 pre-specified strings which may indicate that the activities are "performed" by the company or the company "operates in" a certain activity sector. For example:

- actu-a (-ar, -ando) (em|na)?; act-s (-ing) (in|on)?
- especializad-a (-as, -o, -os) (em|na)?; specializ-ing (-ed) (in|on)?
- dedicada (a|à|ao); dedicated to

In ongoing work, we are continually expanding the above white list. We are also implementing a black list as explained on section 2.4.

Step 4. Identification of sentences containing: "performed" terms and activities
In order to distinguish between activities performed by the company and activities somehow related to it, in this step, economic activities co-occurring with a "*performed*" term within the same sentence are classified as performed activities. Economic activities not classified as "*performed*" are considered somehow related to the company and are ignored (or equivalently considered as "*not performed*").

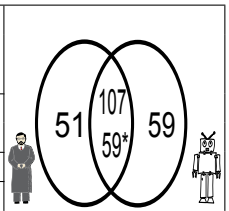
3 Experimental results

The rules used by the system provide an idealized model of how performed economic activities are expressed. Therefore we pose the questions: I) What percentage of the total number of correct performed activities, the system was able to identify (recall)? II) What percentage of the performed activities identified by the system was actually correct (precision)?

For answering these questions, we manually annotated a set of 100 random web pages, each one belonging to a distinct company, with the *economic activities performed* by them. This resulted in a set of 217 performed activities. Running the system on the same dataset resulted in a set of 225 performed activities. The union of the two sets result in 276 unique performed economic activities distributed as shown on Table 3.

As shown on Table 3, 107 of the economic activities returned by the system were totally correct, this mean completely recognised and classified as "*performed*". Hence the recall was 49% (107/217) and the precision was 48% (107/225). Note that in these calculations, the set of overlap activities was considered incorrect. This is because, although all activities in the set overlap are correctly classified as "*performed*", none of them were correctly recognised. The most common error was to miss one or more words as illustrated by the following example, in which the string "[$_{ACTIVITY}$ *Conservação e Restauro de Arte Contemporânea*]" (*Conservation and Restoration of Contemporary Art*) was judged as the correct activity, but the system returned "[ACTIVITY *Conservação e*

Table 3. Set of *performed economic activities* manually annotated and set of *performed economic activities* produced by the system

In both sets (completely recognised and correctly classified as "*performed*")	107	
Only in the manual set (missed by the system)	51	
Only in the system set	59	
Overlap (partially annotated and correctly classified as "*performed*")	59*	

Restauro de Arte]" (*Conservation and Restoration of Art*), missing the word "*contemporânea*" (*contemporary*).

Taking into account what has been said, we computed the *Precision* (P), *Recall* (R), as well their combination into the *F1-measure* according to 3 different criteria: *strict*, *lenient* and *average*. All results were micro-averaged and are shown on Table 4. The *strict* criteria considers all partially correct recognised activities as incorrect. The *lenient* criteria considers all partially correct recognised activities as correct. The *average* criteria allocates a half weight to partially correct recognised activities (i.e. it takes the average of strict and lenient).

Table 4. Experimental results for economic activities

	Strict			**Lenient**			**Average**		
	P	R	F1	P	R	F1	P	R	F1
Micro-average	48%	49%	48%	74%	76%	75%	61%	63%	62%

The difference between strict results and their respective lenient or average results show that a significant proportion of economic activities and sectors identified by the system are partially correct. Those differences may also provide an indication of how challenging finding the exact boundary of an economic activity can be.

3.1 Error Analysis

For understanding the limitations and deficiencies of the system, we performed an error analysis. We analysed:

- The incorrect performed activities returned by the system (errors in the precision).
- The corrected performed activities that the system missed (errors in the recall).

Table 5 (1) summarizes the kinds of incorrect performed economic activities returned by the system. We verified that 50% of the incorrect performed economic activities returned by the system were cases where the economic activity was partially recognised even though correctly classified as performed. For instance, the returned economic activity "*[$_{ACTIVITY}$ transporte de cereais] a garnel*" (*bulk [ACTIVITY transport of cereals]*) is in fact a performed activity, however the string "*a garnel*" (*bulk*) was not recognised. The second causes of errors in 37% of the cases were activities that weren't in fact economic activities. This error can have different causes, the most common was annotating economic activities out of context. For example, in the sentence "*A empresa garante aos seus colaboradores as condições de saúde e segurança no desenvolvimento da sua atividade.*" (The company guarantees its employees, conditions of health and safety in the development of their activity), the system returned "*saúde*" (*health*) and "*segurança*" (*safety*) as economic activities, which is not true in the context of this sentence. These errors are due to the fact that these two strings co-occur in the same sentence with the word "*atividade*" (*activity*). If the system already had a black list of words, as explained on section 2.4, we could introduce the expression "*conditions of*" in it to nullify the effect of the word "*activity*".

Table 5. (1) Incorrect performed economic activities returned by the system. (2) Missed performed economic activities

Incorrect Performed Ec. Activities		
50%	59	Partially recognised, correctly classified as performed
31%	37	Not an economic activity
11%	13	Completely and correctly recognised, incorrectly classified as performed
8%	9	Partially recognised, incorrectly classified as performed
100%	118	

Missed Performed Ec. Activities		
70%	36	Missed economic activity
16%	8	Correctly recognised as activity, incorrectly classified as performed
14%	7	Partially recognised as activity, incorrectly classified as performed
100%	51	

Table 5 (2) summarizes the kinds of corrected *performed economic activities* that the system missed. We verified that the majority of the missed performed activities (70%) were due to missed *economic activities*, this is, economic activities that weren't even recognised. Other sources of failure were due to economic activities correctly or partially recognised but then incorrectly classified as *performed* activity.

Because we were interested in understanding why the system missed 51 of the corrected performed economic activities, we did a deeper analysis. Table 6 summarizes that analysis showing the causes of corrected performed economic

activities that the system missed. The table shows that 27% of the missed performed activities (27%) were caused by strings not inserted in the list (gazetteers) used by the system. These errors are due to the small size of the gazetteers.

Table 6. Causes of the missed performed economic activities

Missed Performed Economic Activities Causes		
27%	14	String not in the gazetteers
24%	12	Not coverage by the rules
22%	11	Implementation mistake
12%	6	Unknown semantic class
6%	3	Not continuous semantic classes
6%	3	Insufficient context
4%	2	Spelling error
100%	51	

4 Related Work

This section is focused on works about information extraction applied to business intelligence.

The computational research aiming at automatically identifying company activities in text, was mentioned on the earlier MUC's (Message Understanding Conferences), involving joint ventures from business news (MUC-5), in both English and Japanese [1]. However, the company activity arises in a context in which it's just one more piece or information among others that should be extracted [8].

In more recent work, and applied to business intelligence, Horacio Saggion et al [11], described the extraction of information for internationalisation applications. Among the information to be extracted by the system are the company name, its main activities, its number of employees, its board of directors, etc. The data are drawn from various types of documents, but also of websites such as Yahoo! Finance, World Bank, CIA Fact Book. The extraction is done by using GATE [4].

Over time, other works have been applied information extraction techniques for business intelligence. As shown below, the extracted information can vary. For example: In [12] the authors seek to identify "management succession". They seek to identify events in which corporate managers left their posts or assumed new ones. To this end, the authors apply a machine learning based approach to automatically identify patterns from annotated text. Starting with a small initial set of patterns proposed by the user, the system is applied incrementally to identify new patterns.

In [6] is presented an application to extract and monitor trends and topics in the field of chemical engineering. The application can monitor resources available on the Internet, such as job advertisements and news. The implemented

prototype uses GATE [4]. The topics to be monitored are those specified by an ontology.

In [3] it is shown how the automatic extraction of data from web sites can be used to obtain information from competitors to support decision-making. The extracted data, which are prices and product information, are then integrated into a business intelligence system. The extraction is done using the Lixto [2], a paid tool that extracts data visually and without any programming knowledge.

In [9], the project MBOI (Matching Business Opportunities on the Internet) is presented a tool for discovering business opportunities on the internet. The purpose of the tool is to help the user decide which tenders should be analyzed. The authors use techniques of information extraction and classification to achieve their goals. While the information extraction techniques are used to obtain data from solicitation notices, the classification techniques are used to classify these competitions under different classification systems, such as the CPV (Common Procurement Vocabulary), SIC (Standard Industrial Classification), NAICS (North American Industry Classification System), FCS (Federal Supply Codes), etc.

In the mentioned works the economic activities is just a piece of information witch should be extracted, among others. Therefore is not clear how the authors addressed the economic activity recognition problem in particular. For the same reason, it was also impossible to compare any results of the different studies.

5 Conclusion and Future Work

We describe a system for the automatic recognition and classification of performed economic activities from text. The approach is based mainly on the idea that an economic activity can be decomposed in semantic classes. Then instead of a huge list of economic activities which is neither practical to use, nor maintain, we can use smaller lists, one for each semantic class, that are easier to maintain and flexible to use. Those lists are then used for recognising economic activities based on syntactic rules. We have empirically tested the validity of the idea and kind of rules used with already very interesting results. We also noted that, although economic activities are a very specific type of information, there still exist many ways of expressing it. The results shows also, how challenging can be to find the exact boundary of an economic activity.

The most immediate and main purpose of the system, is to be used to fill a database on which for each company, we have the economic activities they perform. Before that, we need to expand the current gazetteers and set of rules for handling some hard cases. Until now we have implemented a first set of rules for dealing with the most obvious descriptions of economic activities. These descriptions were sequences of adjacent or almost adjacent strings, belonging to a known semantic class (i.e. ACTION, PRODUCT, SECTOR). As a long-term goal, one would like to investigate the possibility of using an adapted POS-tagger for labelling the semantic classes, similarity to what was done by [5], for labelling postal address components (e.g. street, postal code, etc.).

Acknowledgements

This project was supported by the Portuguese National Strategic Reference Framework — *QREN/Programa Operacional de Factores de Competitividade*, proposal n. 18627 — 06/2010 SI I&DT supported by European Regional Development Fund.

References

1. Proceedings of the 5th conference on Message understanding. In: MUC5 '93: Proceedings of the 5th conference on Message understanding. Association for Computational Linguistics, Baltimore, Maryland (1993)
2. Baumgartner, R., Flesca, S., Gottlob, G.: Visual Web Information Extraction with Lixto. In: The VLDB Journal. vol. 27th, pp. 119–128. Morgan Kaufmann Publishers Inc. (2001)
3. Baumgartner, R., Frölich, O., Gottlob, G., Harz, P., Herzog, M., Lehmann, P., Wien, T.: Web data extraction for business intelligence: the lixto approach. In: In Business. pp. 48–65 (2005)
4. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A framework and graphical development environment for robust NLP tools and applications. In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics. pp. 168–175. Philadelphia, USA (2002)
5. Marques, N.C., E, Calves, S.G.: Applying a Part-of-Speech Tagger to Postal Address Detection on the Web. In: Proceedings of the IV International Conference on Language Resources and Evaluation. pp. 287–290 (2004)
6. Maynard, D., Yankova, M., Kourakis, A., Kokossis, A.: Ontology-based information extraction for market monitoring and technology watch. In: ESWC Workshop End User Apects of the Semantic Web. Heraklion, Crete (2005)
7. McDonald, D.D.: Internal and External Evidence in the Identification and Semantic Categorization of Proper Names. In: Boguraev, B., Pustejovski, J. (eds.) Corpus Processing for Lexical Acquisition, chap. 2, pp. 21–39. MIT Press (1996)
8. Onyshkevych, B., Okurowski, M.E., Carlson, L.: Tasks, domains, and languages. In: Proceedings of the 5th conference on Message understanding - MUC5 '93. pp. 7—-17. Association for Computational Linguistics, Morristown, NJ, USA (1993)
9. Paradis, F., Nie, J.Y., Tajarobi, A.: Discovery of business opportunities on the internet with information extraction. In: Workshop on Multi-Agent Information Retrieval and Recommender Systems (IJCAI). pp. 47–54. Edinburgh, Scotland (2005)
10. Pustejovsky, J.: The Generative Lexicon. Computational Linguistics (4), 409–441
11. Saggion, H., Funk, A., Maynard, D., Bontcheva, K.: Ontology-based Information Extraction for Business Intelligence. In: Aberer, K., Choi, K.S., Noy, N., Allemang, D., Lee, K.I., Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) The Semantic Web. LNCS, vol. 4825, pp. 843–856. Springer (2007)
12. Yangarber, R., Grishman, R., Tapanainen, P., Huttunen, S.: Unsupervised discovery of scenario-level patterns for Information Extraction. In: Proceedings of the sixth conference on Applied natural language processing -. pp. 282–289. Association for Computational Linguistics, Morristown, NJ, USA (2000)