Eksairesis: A Domain-adaptable System for Ontology Building from Unstructured Text

K. L. Kermanidis[1], A. Thanopoulos[2], M. Maragoudakis[3], N. Fakotakis[2]

[1]Department of Informatics
Ionian University, Corfu
7 Pl. Tsirigoti, 49100, Greece
kerman@ionio.gr

[2]Wire Communications Laboratory
Department of Electrical and Computer Engineering
University of Patras, Rio 26500, Greece
{aristom, fakotaki}@wcl.ee.upatras.gr

[3]Department of Information and
Communication Systems Engineering
University of the Aegean, Samos, Greece
mmarag@aegean.gr

Abstract

This paper describes Eksairesis, a system for learning economic domain knowledge automatically from Modern Greek text. The knowledge is in the form of economic terms and the semantic relations that govern them. The entire process in based on the use of minimal language-dependent tools, no external linguistic resources, and merely free, unstructured text. The methodology is thereby easily portable to other domains and other languages. The text is pre-processed with basic morphological annotation, and semantic (named and other) entities are identified using supervised learning techniques. Statistical filtering, i.e. corpora comparison is used to extract domain terms and supervised learning is again employed to detect the semantic relations between pairs of terms. Advanced classification schemata, ensemble learning, and one-sided sampling, are experimented with in order to deal with the noise in the data, which is unavoidable due to the low pre-processing level and the lack of sophisticated resources. An average f-score of 68,5% over all the classes is achieved when learning semantic relations. Bearing in mind the use of minimal resources and the highly automated nature of the process, classification performance is very promising, compared to results reported in previous work.

1. Introduction

This paper describes EKSAIRESIS, a system that has been developed for extracting semantic information from free, unstructured Modern Greek corpora, in order to build an ontology consisting of economic domain terms and the semantic relations linking them together. EKSAIRESIS performs basic morphosyntactic pre-processing, employs statistical filtering to filter out corpus words that do not belong to the domain, and supervised learning to learn the semantic relations between the extracted domain terms. Figure 1 shows a rough overview of the system architecture.

A domain ontology is the tool that enables information retrieval, data mining, intelligent search in a particular thematic domain. Ontologies consist of concepts that are important for communicating domain knowledge. These concepts are structured hierarchically through taxonomic relations. A taxonomy usually includes *hyperonymy-hyponymy* (is-a), and *meronymy* (part-of) relations. Learning taxonomic relations between the concepts that describe a specific domain automatically from corpora is a key step towards ontology engineering. The advent of the semantic web has pushed the construction of concept taxonomies to the top of the list of interests of language processing experts.

A complete ontology may also include further information regarding each concept. The economic domain, especially, is governed by more 'abstract' relations, that capture concept attributes (e.g. *rise* and *drop* are two attributes of the concept *value*, a *stockholder* is an attribute of the concept *company*). Henceforth, this type of relation will be referred to as *attribute* relation.

Unlike most previous approaches, that focus basically on *hyponymy*, in this work, *meronymy* as well as *attribute* relations are also detected. A term pair is governed by an attribute relation if it does not match the typical profile of an *is-a* or a *part-of* relation. All the aforementioned types of relations are henceforth called *taxonomic* in this paper.

One more novel aspect of the described system is that it does not rely on any external resources (e.g. semantic networks, like WordNet, grammars, hierarchically structured corpora, or pre-existing ontologies). Thereby, it is easily applicable to other thematic domains. without any alterations, merely by using a corpus that is relevant to the domain. It should also be noted that the concept hierarchy is built from scratch, instead of trying to extend an already existing ontology.

The lack of sophisticated resources leads unavoidably to the presence of noise in the data. Advanced learning schemata like sampling and ensemble learning (bagging, boosting and stacking) have been employed to deal with the noise and help the learner focus on the useful content data.

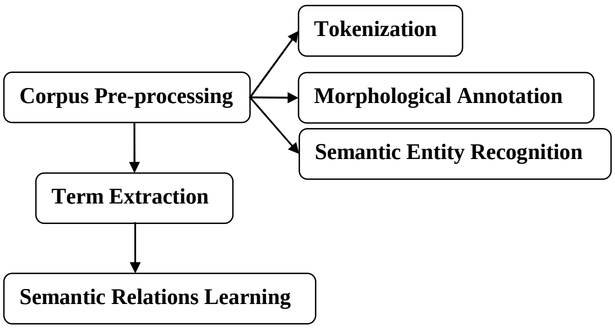

Figure 1: System overview

Noise appears in the form of class imbalance. Positive class instances (instances of the class of interest that needs to be learned) in the data are underrepresented compared to negative instances (null class instances). Class imbalance has been dealt with in previous work in various ways: oversampling of the minority class until it consists of as many examples as the majority class (Japkowicz, 2000), undersampling of the majority class (random or focused), the use of cost-sensitive classifiers (Domingos, 1999), the ROC convex hull method (Provost and Fawcett, 2001). In this work we employ random and focused undersampling.

Another interesting aspect of the present work is the language itself. Modern Greek is a relatively free-word-order language, i.e. the ordering of the constituents of a sentence is not strictly fixed, like it is in English. The rich morphology as well as other typical idiosyncrasies of the language are taken into account throughout the present work. The language-dependent features of the system, however, are not based on the use of hard-to-develop tools or thesauri, allowing the methodology to be easily adaptable to other languages.

The rest of the paper is organized as follows. Section 2 lists previous supervised and unsupervised approaches to taxonomy learning. Section 3 introduces the corpora used in our experiments and their pre-processing, including semantic entity recognition. Section 4 describes the term extraction process, while section 5 presents in detail the taxonomy learning approach. Section 6 discusses some interesting aspects regarding the system, and the paper is completed with some final remarks in section 7.

2. Related Work

Regarding previous approaches to taxonomy learning that employ clustering techniques, (Cimiano et. al., 2004) describe a conceptual clustering method that is based on the Formal Concept Analysis for automatic taxonomy construction from text and compares it to similarity-based clustering (agglomerative and Bi-Section-KMeans clustering). The automatically generated ontology is compared against a hand-crafted gold standard ontology for the tourism domain and a maximum lexical recall of 44.6% is reported.

Other clustering approaches are (Faure and Nedellec, 1998; Pereira et al., 1993). The former uses a parsed text and utilizes iterative clustering to form new concept graphs. The latter also makes use of verb-object dependencies, relative frequencies and relative entropy as similarity metrics for clustering.

Pekar and Staab (2002) take advantage of a taxonomic thesaurus (a tourism-domain ontology) to improve the accuracy of classifying new words into its classes. Their classification algorithm is an extension of the kNN method, which takes into account the taxonomic similarity between nearest neighbors. They report a maximum overall accuracy of 43.2%.

Lendvai (2005) identifies taxonomic relations between two sections of a medical document using memory-based learning. Binary vectors represent overlap between the two sections, and the tests are run on parts of two Dutch medical encyclopedias. A best overall accuracy value of 88% is reported.

Witschel (2005) proposes a methodology for extending lexical taxonomies by first identifying domain-specific concepts, then calculating semantic similarities between concepts, and finally using decision trees to insert new concepts to the right position in the taxonomy tree.

Navigli and Velardi (2004) interpret semantically the set of complex terms that they extract, based on simple string inclusion. They use a variety of external resources in order to generate a semantic graph of senses.

Another approach that makes use of external hierarchically structured textual resources is (Makagonov et al., 2005). An already existing hierarchical structure of technical documents is mapped to the structure of a domain-specific technical ontology. Words are clustered into concepts, and concepts into topics. The ontology is evaluated against the structure of existing textbooks in the given domain.

Maedche and Volz (2001) make use of clustering, as well as pattern-based (regular expressions) approaches in order to extract taxonomies from domain-specific German texts.

Degeratu and Hatzivassiloglou (2004) also make use of syntactic patterns to extract hierarchical relations, and measure the dissimilarity between the attributes of the terms using the Lance and Williams coefficient. They evaluate their methodology on a collection of forms provided by the state agencies and report a precision value of 73% and 85% for *is-a* and attributive relations respectively.

As can be seen, most previous approaches rely on the use of external resources, semantic networks, grammars, pre-existing ontologies, hierarchically structured corpora, and focus mainly on *is-a* relations. In this work, we use merely an unstructured corpus and attempt to learn a wider range of relation types.

3. Corpora and Pre-processing

The corpora that have been used are two document collections: one domain-specific (economic), and one balanced. The use of the two corpora allows for Corpora Comparison for term extraction. The domain-specific corpus (Kermanidis et al., 2002) is a collection of economic texts of approximately 5M words of varying genre, which has been automatically annotated from the ground up to the level of phrase chunking. The balanced corpus (Hatzigeorgiu et al., 2000; Partners, 1986) is a collection of 300,000 words, manually annotated with morphological information. Phrase chunking was performed on both corpora using the tool described in (Stamatatos et al., 2000).

3.1. Semantic Entity Recognition

Identifying names is an important step towards the automatic extraction of domain terms, as each of these names may constitute a candidate term. Each token in the domain-specific corpus constitutes a candidate semantic entity. The semantic entities in the present work are not limited to named entities only, such as names of organizations, persons and locations, but they belong to one of 31 distinct classes: names of organizations, persons and locations, names of stocks and bonds, names of newspapers and magazines, quantitative units, amounts, values, percentages, temporal expressions. All these entity classes are important for data mining tasks. Recognition of these entities is viewed as a two-task experiment: The first task is to detect the boundaries of the entities. The second is to assign a semantic label to each entity.

The Modern Greek language has certain properties that are significant for the present task. First, it is highly inflectional. The case (nominative, accusative, genitive) of nouns, adjectives or articles affects semantic labeling. For example, the genitive case may denote possession, quantity, quality, origin, division, etc., as is shown in the fol-lowing examples:

```
Η τιμή ανήλθε στο ποσό των 12.33 €.
The[NOM] price[NOM] reached the[ACC]
value[ACC] the[GEN] 12.33 €.
(The price reached the value of 12.33 €.)
```

```
Η Τράπεζα της Ελλάδος
The[NOM] Bank[NOM] the[GEN] Greece[GEN]
(The Bank of Greece)
```

Supervised learning techniques have been employed to learn the boundaries and the labels of the entities. Each candidate entity is represented by a feature-value vector. The features forming the vector are:

1. The token lemma.
2. The part-of-speech category of the token.
3. The morphological tag of the token. The morphological tag is a string of 3 characters encoding the case, number, and gender of the token.
4. The case tag of the token.
5. A Boolean feature encoding whether the first letter of the token is capitalized.

Context information is often decisive when trying to detect a semantic entity. In the following example, the verb *ανέρχομαι* (to reach), is a strong indicator that the entity next to it is an amount/value, because this verb is typically used in Modern Greek to express 'reaching a value'

```
Οι μετοχές ανήλθαν στις 500.
The stocks reached the 500.
(The number of stocks reached 500.)
```

Context information was included in the feature-value vector, by taking into account a window of tokens preceding and following the candidate entity. Various experiments have been run to determine the optimal window size, which depends on the entities to be learned (Kermanidis, 2007b).

A total of 40,000 tokens were manually tagged with their class value. An interesting observation is the imbalance between the populations of the positive instances (entities) in the dataset (that form only 15% of the total number of instances) and the negative instances (non-entities). This imbalance has serious consequences on classification accuracy of the instances of the minority classes. By randomly removing negative examples (undersampling), so that their number reaches that of the positive examples, the imbalance is attacked and the results prove that classification accuracy of the positive instances improves.

The results improve further with two-phase learning: the learner is first trained on the training data and used to classify new, unseen instances. In the second stage, the classification predictions of the first stage are added to the instance vector as extra features to force the classifier to learn from its mistakes. The augmented vector is fed to a higher-level learner, which makes a final prediction.

Instance-based learning (1-NN) was selected to classify the candidate semantic entities. 1-NN was chosen because, by storing all examples in memory, it is able to deal competently with exceptions and low-frequency events, which are important in language learning tasks, and are ignored by other learning algorithms. Table 1 shows the average f-score over all classes for context window (-1, +1). These results were obtained using 10-fold cross validation. The beneficial impact of learning in two phases and undersampling is evident.

Method	Avg f-score
Initial dataset	0.55
Two-phase learning	0.73
Two-phase learning + Undersampling	0.74

Table 1: Semantic entity recognition results

'Straightforward' semantic entity types, i.e. types that are introduced or followed by a limited number of characteristic words, acronyms, abbreviations or symbols (like monetary amounts, percentages, company names), are easier to learn than the rest. This holds, even if the type appearance in the corpus is sparse, like in the case of newspaper/journal names. Their f-score reaches at least 94,4%, although they appear in the data set with a frequency of 0.14%. Their normally being introduced by the word *εφημερίδα* or *περιοδικό* helps identify them accurately.

4. Extracting Economic Terms

The next step of the procedure is the automatic extraction of economic terms, following the methodology described in (Thanopoulos et al., 2006). Corpora comparison was employed for the extraction of economic terms. Corpora comparison detects the difference in statistical behavior that a term presents in a balanced and in a domain-specific corpus.

Noun and prepositional phrases of the two corpora are selected to constitute candidate terms, as only these phrase types are likely to contain terms. The occurrences of words and multi-word units (n-grams), pure as well as nested, are counted. Longer candidate terms are split into smaller units (tri-grams into bi-grams and uni-grams, bi-grams into uni-grams). Due to the relative freedom in the word ordering in Modern Greek sentences, bi-gram A B (A and B being the two lemmata forming the bi-gram) is considered to be identical to bi-gram B A, if the bi-gram is not a semantic entity. Their joint count in the corpora is calculated and taken into account. The resulting uni-grams and bi-grams are the candidate terms. The candidate term counts in the corpora are then used in statistical filters.

Statistical filtering is performed in two stages: First the relative frequencies are calculated for each candidate term. Then, for those candidate terms that present a relative frequency value greater than 1, the Log Likelihood ratio (LLR) is calculated. The LLR metric detects how surprising (or not) it is for a candidate term to appear in the domain-specific or in the balanced corpus (compared to its expected appearance count), and therefore constitute an economic domain term (or not).

Rank	word	translation	Cw(D)	Cw(B)	RFw	LLR
1	εταιρία	company	5396	0	1845,9	852,0
2	δρχ	drachma	3003	1	342,5	465,5
3	μετοχή	stock	2827	6	74,4	414,0
4	αγορά	buy	2330	33	11,9	257,2
5	αύξηση	growth, rise	2746	66	7,1	247,6
6	κέρδος	profit	1820	15	20,1	228,2
7	τράπεζα	bank	1367	11	20,3	171,8
8	επιχείρηση	enterprise	1969	56	6,0	162,1
9	κεφάλαιο	capital	1325	14	15,6	157,3
10	σημαντικός	important	1872	56	5,7	149,3
11	πώληση	sell	1203	11	17,9	147,3
12	προϊόν	product	1282	16	13,3	146,0
13	όμιλος	(company) group	1036	5	32,2	140,0
14	Α.Ε.	INC	820	0	280,7	126,4
15	μετοχικός	stocking	790	2	54,1	112,8
16	τιμή	price	1722	70	4,2	110,9
17	επιτόκιο	interest (financ.)	821	4	31,2	110,0
18	υψηλός	high (old form)	711	0	243,4	109,2
19	κόστος	cost	1031	19	9,0	103,4
20	κλάδος	branch	833	7	19,0	103,2

Table 2: The 20 most highly ranked economic terms

Table 2 shows the relative frequency (RFw) and LLR scores of the 20 most highly ranked economic terms, ordered by their LLR value. Cw(D) and Cw(B) are the term counts in the domain-specific and the balanced corpus respectively. An interesting term is “υψηλός”, the ancient Greek form for “high”, used today almost exclusively in the context of the degree of performance, growth, rise, profit, cost, drop (i.e. the appropriate form in economic context), as opposed to its modern form “ψηλός”, which is used in the concept of the degree of actual height.

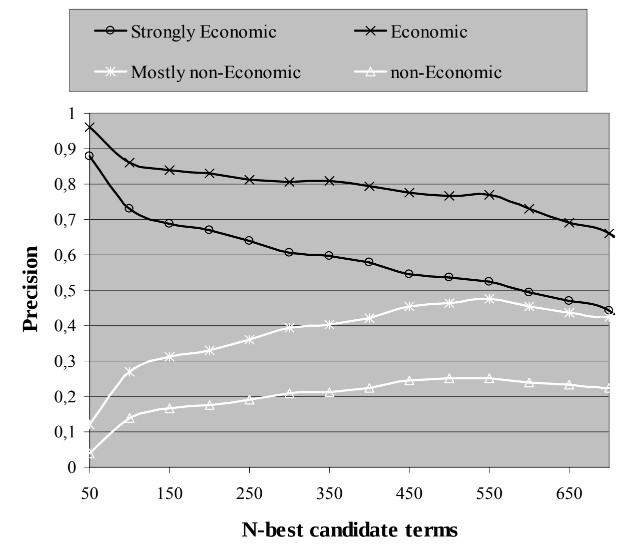

Figure 2: Precision (y-axis) for the N-best candidate terms (x-axis) that appear in both corpora.

As can be seen in Figure 2, the term extraction methodology reaches a precision of 82% for the 200 N-best candidate terms. In this figure, *strongly economic* are terms that are characteristic of the domain and necessary for understanding domain texts. *Economic* are terms that function as economic within a context of this domain, but may also have a different meaning outside this domain. *Mostly non-economic* are words that are connected to the specific domain only indirectly, or more general terms that normally appear outside the economic domain, but may carry an economic sense in certain limited cases. *Non-economic* are terms that never appear in an economic sense or can be related to the domain in any way.

5. Learning Semantic Relations

After the term extraction process, the 250 most highly ranked terms (according to the LLR metric) were selected, and each one was paired with the rest. Semantic information regarding each individual term, as well as each term pair is represented through the set of attributes described in the following paragraphs.

It is an axiom in semantics that the sense of a term is strongly linked to the context it appears in. To this end, a context window of two words preceding and two words following the term for every occurrence of the term in the corpus is formed. All non-content words (prepositions, articles, pronouns, conjunctions, etc.) are disregarded, while acronyms, abbreviations, and certain symbols (e.g. %, €) are not, because of their importance for determining the semantic profile of the term in this particular domain. Bi-grams (pairs of the term with each word within the con-text window) are generated and their frequency is recorded. The ten words that present the highest bi-gram frequency scores are chosen to form the *semantic context vector* of the term.

The *semantic similarity* of the two terms forming each pair is estimated using their semantic context vectors. The smaller the distance between the context vectors, the more similar the terms' semantics. The value of semantic similarity is an integer with a value ranging from 0 to 10, which denotes the number of common words two context vectors share. It is inspired by the Dice coefficient (Manning and Schuetze, 1999).

Another important semantic feature that is taken into account is how 'diverse' the semantic properties of a term are, i.e. the number of other terms that a term shares semantic properties with. We estimate the *semantic diversity* of a term by calculating the percentage of the total number of terms whose semantic similarity with the focus term (one of the two terms whose taxonomic relation is to be determined) is at least 1.

The syntactic patterns that govern the co occurrence of two terms, is significant for extracting taxonomic information. For languages with a strict sentence structure, like English, such patterns are easier to detect (Hearst, 1992), and their impact on taxonomy learning more straightforward. Modern Greek presents a larger degree of freedom in the ordering of the constituents of a sentence, due to its rich morphology and its complex declination system. This freedom makes it difficult to detect syntactic patterns, and, even if they are detected, their contribution to our task is not that easily observable.

There are, however, two Modern Greek syntactic schemata prove very useful for learning taxonomies. We name them the *attributive modification schema* and the *genitive modification schema*. The first, known in many languages, is the pattern where (usually) an adjective (A) modifies the following noun (N). The second is typical for Modern Greek, and it is formed by two nominal expressions, one of which (usually following the other) appears in the genitive case (N-GEN) and modifies the preceding nominal, denoting possession, property, origin, quantity, quality. The following phrases show examples of the first (example 1) and the second (example 2) schemata respectively.

```
(1) το     μετοχικό[A]    κεφάλαιο[N]
    the    stock          capital

(2) η   κατάθεση[N]    επιταγής[N-GEN]
    the deposit        check
    (the deposit of the check)
```

The syntactic and semantic information described has again been encoded in a set of attributes that form a feature-value vector for each pair of terms. The features in the vector are:

- the terms' lemmata
- the terms' frequencies
- the terms' part-of-speech tags
- the pair's semantic similarity value
- the terms' semantic diversity values
- the number of times the two terms co-occur in one of the two syntactic schemata in one of the following four relative positions (the underscore denotes an intervening word):

  1. term1 term2
  2. term2 term1
  3. term1 _ term2
  4. term2 _ term1

The semantic relations of a total of 6000 term pairs were manually annotated by economy and finance experts with one of the four class label values: *is-a*, *part-of*, *attribute* relation and no relation (*null*). 9% of the term pairs belong to the *is-a* class, 17% to the attribute class and only 0.5% to the *part-of* class. The *is-a* and *part-of* classes are significantly underrepresented, and this class imbalance has been dealt with using one-sided sampling (Kermanidis and Fakotakis, 2007). One–sided sampling removes from the dataset negative instances that are redundant, noisy or borderline (close to the boundary line that separates the positive from the negative instances). One-sided sampling (OSS) has been employed in several domains (Kubat and Matwin, 1997; Laurikkala, 2001; Lewis and Gale, 1994), and generally leads to better classification performance than oversampling, and it avoids the problem of arbitrarily assigning initial costs to instances.

Experiments were run using the C4.5 decision tree learner (Quinlan, 1993), as a stand-alone learner, as well as a base learner for boosting (Freund and Schapire, 1997). Decision trees were chosen because of their high representational power, which is very significant for understanding the impact of each feature on the classification accuracy, and because of the knowledge that can be extracted from the resulting tree itself. Support vector machines (SVM) were also experimented with, as they constitute a methodology that copes well with the sparse data problem, and also with noise in the data. A first degree polynomial kernel function was selected and the Sequential Minimal Optimization algorithm (Platt, 1998) was chosen to train the support vector classifier. A Naïve Bayes learner and the 1-NN instance based-learner (IB1) were also experimented with as baseline reference.

The f-scores achieved with every stand-alone classifier, for every class, are shown in Table 3. The evaluation was performed again using 10-fold cross validation. The poor results for the *part-of* relation are attributed mainly to its extremely rare occurrence in the data. The economic domain is more "abstract" and is governed to a large extent by other relation types.

To overcome this problem of performance instability among the various classifiers, the application of ensemble learning (Opitz and Maclin, 1997) is proposed. The combination of various disagreeing classifiers leads to a resulting classifier with better overall predictions (Dietterich, 2002). Experiments have been conducted using the aforementioned classifiers in various combination schemes using bagging, boosting (as mentioned previously) and stacking (Kermanidis, 2007a).

	C4.5	IB1	Naïve Bayes	SVM
Is-a	0.808	0.694	0.419	0.728
Part-of	0.4	0	0	0
Attribute	0.769	0.765	0.77	0.788
Null	0.938	0.904	0.892	0.907

Table 3: Class f-score for various classifiers.

Bagging entails the random partitioning of the dataset in equally sized subsets (bags) using resampling. Each subset trains the same base classifier and produces a classification model (hypothesis). The class of every new test instance is predicted by every model, and the class label with the majority vote is assigned to the test instance.

Unlike bagging, where the models are created separately, boosting works iteratively, i.e. each new model is influenced by the performance of those built previously (Freund and Schapire, 1996). In other words, new models are forced, by appropriate weighting, to focus on instances that have been handled incorrectly by older ones.

Finally, stacking usually combines the models created by different base classifiers, unlike bagging and stacking where all base models are constructed by the same classifier (Dietterich, 2002). After constructing the different base models, a new instance is fed into them, and each model predicts a class label. These predictions form the input to another, higher-level classifier (the so-called meta-learner), that combines them into a final prediction.

	Boosting	**Bagging**	**Stacking**	**OSS**
Is-a	0.772	0.856	0.853	0.789
Part-of	0.286	0	0	0.33
Attribute	0.762	0.809	0.835	0.794
Null	0.922	0.962	0.957	0.927

Table 4: Results for learning taxonomic relations.

Table 4 shows the best achieved f-scores. Bagging and boosting achieved the best results using C4.5 as their base learner. For stacking, the best classifier combination was formed by the following base learners: the IB1 instance based-learner, the C4.5 decision tree learner, the Naïve Bayes learner, the Bayes Network classifier and the SVM classifier, and the latter as meta-learner. However, even a simple lazy meta-learner (IB1) reaches an f-score higher than 81% for all three classes (except the *part-of* class). This is attributed to the predictive power of the combination of base learners. In other words, the sophisticated base learners do all the hard work, deal with the difficult cases, and the remaining work for the meta-learner is simple.

Comparing Tables 3 and 4, the positive impact of combining multiple classifiers into a single prediction scheme becomes apparent. Mistakes made by one single classifier are amended through the iterative process and the majority voting in bagging, through instance weighting, according to how difficult an instance is to predict, in boosting, and through combining the strengths of several distinct classifiers in stacking.

6. Discussion

The type of semantic relation between two terms is not always straightforward, especially in a domain as ‘abstract’ as the economic domain. For example, the pair μέτοχος-συνέλευση (stockholder-board) was tagged as part-of by the experts (the stockholders form the body of the board and are therefore part of it) and it was predicted as attribute by the learner. Due to the ‘collective’ meaning of the attribute class (it includes all relation types except *is-a* and *part-of*), it is not clear how ‘big’ the learner’s mistake is. A more fine-grained distinction between the types of attribute relations is a challenging future research direction, providing information that is very useful for data mining applications in the domain.

The extracted relations are useful in many ways. They form a generic semantic thesaurus that can be further used in several applications. First, the knowledge is important for economy/finance experts for a better understanding and usage of domain concepts. Moreover, the thesaurus facilitates intelligent search. Looking for semantically related terms improves the quality of the search results. The same holds for information retrieval and data mining applications. Intelligent question/answering systems that take into account terms that are semantically related to the terms appearing in queries return information that is more relevant, more accurate and more complete.

The next research step for EKSAIRESIS will be the organization of the extracted terms and the relations governing them into a complete hierarchical ontological structure, so that the acquired knowledge can be utilized by ontology tools and ontology-based applications.

7. Conclusion

In the present work we have described a system for learning semantic relation between economic domain terms. The semantic knowledge is extracted automatically from free, unstructured Modern Greek corpora. The methodology makes use of no additional external resources, allowing the process to be easily applicable to other thematic domains. Furthermore, the use of no language-dependent tools or thesauri (except for the corpora), makes the system easily portable to other languages.

The lack of sophisticated resources and the low pre-processing level inevitably allow for a significant amount of noise to enter the data. Noise appears in the form of imbalance in the distribution between the instances of interest and the null instances in our learning experiments. This imbalance is dealt using sophisticated classification schemata, like sampling and ensemble learning. Bearing in mind the use of minimal resources and the highly automated nature of the process, classification performance is very promising, compared to results reported in previous work

8. Acknowledgements

We thank the European Social Fund (ESF), Operational Program for Educational and Vocational Training II (EPEAEK II), and particularly the Program PYTHAGORAS II, for funding the above work.

9. References

Cimiano, P., Hotho, A., Staab, S. (2004). Comparing conceptual, divisive and agglomerative clustering for learning taxonomies from text. *Proceedings of the European Conference on Artificial Intelligence*, Spain.

Degeratu, M., Hatzivassiloglou, V. (2004). An automatic model for constructing domain-Specific ontology resources. *International Conference on Language Resources and Evaluation*, Portugal, pp. 2001-2004.

Dietterich, T. (2002). *Ensemble learning. The handbook of brain theory and neural networks.* Second Edition. Cambridge MA, The MIT Press.

Domingos, P. (1999). Metacost: A general method for making classifiers cost-sensitive. *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, San Diego, CA, pp. 155-164.

Faure, D., Nedellec., C. (1998). A Corpus-based conceptual clustering method for verb frames and ontology. *Proceedings of the LREC Workshop on Adapting Lexical and Corpus Resources to Sublanguages and Applications*, Granada, Spain.

Freund, Y., Schapire, R. E. (1996). Experiments with a new boosting algorithm. *Proceedings of the International Conference on Machine Learning*, San Francisco, pp. 148-156.

Hatzigeorgiu, N., et al. (2000). Design and implementation of the online ILSP Greek corpus. *2nd International Conference on Language Resources and Evaluation*, Athens, Greece, pp. 1737−1742.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. *Proceedings of the International Conference on Computational Linguistics*, Nantes, France, pp. 539-545.

Japkowicz, N. (2000). The class imbalance problem: significance and strategies. *Proceedings of the International Conference on Artificial Intelligence*, Las Vegas.

Kermanidis, K., Fakotakis, N., Kokkinakis, G. (2002). DELOS: An automatically tagged economic corpus for Modern Greek. *3rd International Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria, Spain, pp. 93-100.

Kermanidis, K. (2007a). Ensemble learning of economic taxonomy relations from Modern Greek corpora. *2nd International Conference on Metadata and Semantics Research,* Corfu, Greece.

Kermanidis K. (2007b). Identifying boundaries and semantic labels of economic entities using stacking and re-sampling. *Internaational Workshop on Natural Language Processing and Cognitive Science*, Madeira, Portugal.

Kermanidis, K., Fakotakis, N. (2007). One-sided sampling for learning taxonomic relations in the Modern Greek economic domain. *IEEE International Conference on Tools with Artificial Intelligence,* Patras, Greece.

Kubat, M., Matwin, S. (1997). Addressing the curse of imbalanced training sets. *Proceedings of the International Conference on Machine Learning*, pp. 179- 186.

Laurikkala, J. (2001). Improving identification of difficult small classes by balancing class distribution. *Proceedings of the 8th Conference on Artificial Intelligence in Medicine in Europe*, Cascais, Portugal, pp. 63-66.

Lendvai, P. (2005). Conceptual taxonomy identification in medical documents. *International Workshop on Knowledge Discovery and Ontologies*, Porto, Portugal, pp. 31-38.

Lewis, D., Gale, W. (1994). Training text classifiers by uncertainty sampling. *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, Dublin, pp. 3-12.

Maedche, A., Volz, R. (2001). The ontology extraction and maintenance framework Text-To-Onto. *Workshop on Integrating Data Mining and Knowledge Mining*, San Jose, California.

Makagonov, P., Figueroa, A. R., Sboychakov, K., Gelbukh, A. (2005). Learning a domain ontology from hierarchically structured texts. *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany.

Manning, C., Schuetze., H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Navigli, R., Velardi, P. (2004). Learning domain ontologies from document warehouses and dedicated websites. *Computational Linguistics,* 50(2). MIT Press.

Partners of ESPRIT-291/860. (1986). Unification of the word classes of the ESPRIT Project 860. Internal Report BU-WKL-0376.

Opitz, D., Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research* (11), pp. 169-198.

Pekar, V., Staab. S. (2002). Taxonomy learning –factoring the structure of a taxonomy into a semantic classification decision. *Proceedings of the International Conference on Computational Linguistics*, Taipei, Taiwan.

Pereira, F., Tishby, N., Lee, L. (1993). Distributional clustering of English words. *Proceedings of the 31st*

Annual Meeting of the Association for Computational Linguistics.

Platt, J. (1998). Fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods - support vector learning*, B. Schoelkopf, C. Burges, and A. Smola, eds. MIT Press.

Provost, F. and Fawcett, T. (2001). Robust classification for imprecise environments. *Machine Learning*, 42(3), pp. 203-231.

Quinlan, R. (1993). *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers, San Mateo, CA.

Schapire, R. E., Rochery, M., Rahim, M., Gupta, N. (2002). Incorporating prior knowledge into boosting. *Proceedings of the Nineteenth International Conference on Machine Learning*.

Stamatatos, E., Fakotakis, N. and Kokkinakis, G.. (2000). A practical chunker for unrestricted text. *Proceedings of the Conference on Natural Language Processing*, Patras, Greece, pp. 139̃-150.

Thanopoulos, A., Kermanidis, K., Fakotakis, N. (2006). Challenges in extracting terminology from Modern Greek texts. *Workshop on Text-based Information Retrieval*, Italy.

Witschel, H. F. (2005). Using decision trees and text mining techniques for extending taxonomies. *Proceedings of the Workshop on Learning and Extending Lexical Ontologies by Using Machine Learning Methods*