Identifying Boundaries and Semantic Labels of Economic Entities Using Stacking and Re-sampling

Katia Lida Kermanidis

Artificial Intelligence Group, Department of Electrical and Computer Engineering
University of Patras, Rio 26500, Greece

Abstract. Semantic entities of the economic domain are detected and labeled in free Modern Greek text using Instance-based learning in two phases (stacking) to force the classifier to learn from its mistakes, and random undersampling of the majority class to improve classification accuracy of the instances of the minority classes. By not making use of any external sources (gazetteers etc), and limited linguistic information for pre-processing, a mean f-score value of 73.3% for the minority classes is achieved.

1 Introduction

The tagging of semantic entities in written text is an important subtask for information retrieval and data mining and refers to the task of identifying the entities and assigning them to the appropriate semantic category.

One major subclass of semantic entities are named entities (such as names of persons, organizations, locations etc.). Automatic named entity recognition (NER) has been attracting the interest of numerous researchers during the last years. Hendrickx and Van den Bosch in [3] employ manually tagged and chunked English and German datasets, and use memory-based learning to learn new named entities that belong to four categories. They perform iterative deepening to optimize their algorithmic parameter and feature selection, and extend the learning strategy by adding seed list (gazetteer) information, by performing stacking and by making use of unannotated data. They report an average f-score on all four categories of 78.20% on the English test set. Another approach that makes use of external gazetteers is described in [1], where a Hidden Markov Model and Semi-Markov Model is applied to the CoNNL 2003 dataset. The authors report a mean f-score of 90%. Multiple stacking is also employed in [10] on Spanish and Dutch data and the authors report 71.49% and 60.93% mean f-score respectively. The work in [9] focuses on the Natural History domain. They employ a Dutch zoological database to learn three different named-entity classes, and use the contents of specific fields of the database to bootstrap the named entity tagger. In order to learn new entities they, too, train a memory-based learner. Their reported average f-measure reaches 68.65% for all three entity classes. Other approaches ([7], [11]) utilize combinations of classifiers in order to tag new named entities by ensemble learning.

Lida Kermanidis K. (2007).
Identifying Boundaries and Semantic Labels of Economic Entities Using Stacking and Re-sampling.
In *Proceedings of the 4th International Workshop on Natural Language Processing and Cognitive Science*, pages 149-158
DOI: 10.5220/0002414101490158

This paper describes the automatic recognition of semantic entities related to the economic domain in Modern Greek free text. Unlike in previous approaches to NER, the semantic entities in the present work are not limited to named entities only, such as names of organizations, persons and locations. First, they also cover names of stocks and bonds, as well as names of newspapers (due to the newswire genre of the corpus). Furthermore, there are other semantic types that are important for economic information retrieval, like quantitative units (e.g. denoting stock and fund quantities, monetary amounts, stock values), percentages etc. Temporal words and expressions are also identified due to their importance for data mining tasks.

This information appears in free text in either one-word, or multi-word expressions. The present work views semantic entity recognition as a two-task experiment: The first task is to detect the boundaries (the beginning and the end) of these expressions. The second is to assign a semantic label to each of them.

The corpus used in the experiments is automatically tagged with part-of-speech (pos), basic token type information (whether a token is a number, a symbol, an abbreviation, an acronym etc.) and elementary morphological information (case, number and gender). This information is represented by a set of features (described in detail in section 3.2) that form instance vectors. Context information is also taken into account for recognizing new entities, as the tokens surrounding the candidate entity often determine the classification outcome.

Supervised learning techniques have been employed to learn the boundaries and the labels of the entities. Learning is performed in two stages: 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.

Another aspect of the present work that differentiates it from previous approaches is the attempt to deal with the class imbalance problem. As every sentence token is considered a candidate semantic entity, the class of negative instances (instances that do not represent an entity) is highly over-represented in the dataset compared to the positive classes (instances that do represent an entity). This imbalance has serious consequences on classification accuracy of the instances of the minority classes. Random under-sampling of the majority class instances is applied to balance the dataset and improve classification performance.

2 Modern Greek

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 following examples:

Η τιμή ανήλθεστο 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

As can be induced from these examples, another important property is the agreement of morphological features (case, person, gender and number values) between consecutive words. The borders of the agreement define the borders of basic nominal chunks.

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.

3 Data

The experiments described in this paper were run on free Modern Greek text of economic domain. This section describes the corpus, as well as the extracted feature set.

3.1 Delos

The DELOS Corpus ([4]) is a collection of economic domain texts of approximately five million words and of varying genre. It has been automatically annotated from the ground up. Lemmatization and morphological tagging on DELOS was performed by the analyzer described in [8]. Regarding the morphological information that is crucial for the present task, tagging includes assigning part-of-speech (pos) categories to words, assigning case, number and gender tags, detecting acronyms, abbreviations, numbers and symbols. Accuracy in part-of-speech and case tagging reaches 98% and 94% respectively.

DELOS is a collection of newspaper and journal articles. More specifically, the collection consists of texts taken from the financial newspaper EXPRESS, reports from the Foundation for Economic and Industrial Research, research papers from the Athens University of Economics and several reports from the Bank of Greece. The documents are of varying genre like press reportage, news, articles, interviews and scientific studies and cover all the basic areas of the economic domain, i.e. microeconomics, macroeconomics, international economics, finance, business administration, economic history, economic law, public economics etc. Therefore, it presents a richness in vocabulary, in linguistic structure, in the use of idiomatic expressions and colloquialisms, which is not encountered in the highly domain- and language-restricted texts used normally for named entity recognition (e.g. medical records, technical articles, tourist site descriptions).

The following table presents some statistical data regarding the composition of the corpus. Residuals include transliterated words (foreign words written in the Greek alphabet) and interjections.

Table 1. Statistical data on Delos.

Phrases	Noun	36,6%
	Verb	30,9%
	Prepositional	27%
	Adverbial	5,5%
Word tokens	Words	84,1%
	Punctuation marks	8,9%
	Abbreviations/Acronyms	3,3%
	Numbers	2,9%
	Other Symbols	0,8%
Words	In Greek alphabet	98,2%
	In Latin alphabet	1,8%
Words in Greek	Nouns	29,9%
	Verbs	10,2%
	Adjectives	10,6%
	Pronouns	3%
	Articles	15,2%
	Adverbs	5,8%
	Numerals	1,5%
	Conjunctions	6%
	Particles	1,6%
	Prepositions	9,2%
	Residuals	7%

3.2 Feature Set

Each token in the corpus constitutes a candidate semantic entity. Each candidate entity is represented by a feature-value vector. The features forming the vector are:

1. The token lemma. In the case where automatic lemmatization was not able to produce the token lemma, the token itself is the value of this feature.
2. The pos category of the token. The values of this feature appear in table 2.
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, if it is nominal (noun, adjective or article).
4. The case tag of the token. The case tag is one of three characters denoting the token case.
5. Capitalization. A Boolean feature encoding whether the first letter of the token is capitalized or not.

Table 2. Values of the pos feature.

Tag	Description
N	Noun
V	Verb
A	Adjective
P	Pronoun
T	Article
N	Numeral
C	Conjunction
R	Adverb
S	Preposition
F	Punctuation mark
U	Particle
Xa	Acronym
Xb	Abbreviation

For each candidate entity, context information was included in the feature-value vector, by taking into account the two tokens preceding and the two tokens following it. Each of these tokens was represented in the vector by the five features described above. As a result, a total of 25 (5x5) features are used to form the instance vectors.

The class label assigns a semantic tag to each candidate token. These tags represent the entity boundaries (whether the candidate token is the start, the end or inside an entity) as well as the semantic identity of the token. A total of 40,000 tokens were manually tagged with their class value. Table 3 shows the various values of the class feature, as well as their frequency among the total number of tokens.

Unlike most previous approaches that focus on labeling three or four semantic categories of named entities, the present work deals with a total of 30 class values plus the non-entity (NULL) value, as can be seen in the previous table.

Another important piece of information provided disclosed by the previous table is the imbalance between the populations of the positive instances (entities) in the data-set, that form only 15% of the total number of instances, and the negative instances (non-entities). This imbalance leads to serious classification problems when trying to classify instances that belong to one of the minority classes ([5]). By randomly removing negative examples, so that their number reaches that of the positive examples ([6]), the imbalance is attacked and the results prove that classification accuracy of the positive instances improves considerably.

4 Experimental Setup and Results

Instance-based learning (IB1) was the algorithm selected to classify the candidate semantic entities. Ib1 was chosen because, due to storing all examples in memory, it is able to deal competently with exceptions and low-frequency events, which are important in language learning tasks ([2]), and are ignored by other learning algorithms.

Several experiments were conducted for determining the optimal context window size of the candidate entities. Sizes (-2, +2) - two tokens preceding and two following the candidate entity - and (-1, +1) - one token preceding and one following the candidate entity - were experimented with, and comparative performance results were obtained. When decreasing the size from (-2, +2) to (-1, +1), the number of features forming the instance vectors drops from 25 to 15. The results are shown in table 4.

Another set of experiments focused on comparing classification in one stage and in two stages. In the first stage, the Instance-based learner predicts the class labels of the test instances. The results, as noted previously, are presented in table 3. In the second stage, the predictions of the first phase are added to the set of features that are described in section 3.2. The total number of features in the second stage, when experimenting with the (-2, +2) context window, is 30. The results of learning in two stages with window size (-1, +1) are shown in the first column of table 5.

Table 3. Values of the class label.

Tag	Description	Percentage
AE	Start of company/organization/bank name	1.4%
ME	Middle of company/organization/bank name	0.74%
TE	End of company/organization/bank name	1.4%
E	Company/organization/bank 1-word name	1.1%
AP	Start of monetary amount/price/value	0.88%
MP	Middle of monetary amount/price/value	0.63%
TP	End of monetary amount/price/value	0.88%
AAM	Start of number of stocks/bonds	0.3%
MAM	Middle of number of stocks/bonds	0.42%
TAM	End of number of stocks/bonds	0.3%
AT	Start of percentage value	0.73%
MT	Middle of percentage value	0.08%
TT	End of percentage value	0.73%
AX	Start of temporal expression	1%
MX	Middle of temporal expression	0.75%
TX	End of temporal expression	1%
X	1-word temporal expression	0.55%
AO	Start of stock/bond name	0.16%
MO	Middle of stock/bond name	0.17%
TO	End of stock/bond name	0.16%
ON	1-word stock/bond name	0.05%
AL	Start of location name	0.21%
ML	Middle of location name	0.48%
TL	End of location name	0.21%
L	1-word location name	0.33%
F	1-word newspaper/journal name	0.14%
AN	Start of person name	0.18%
MN	Middle of person name	0.02%
TN	End of person name	0.18%
N	1-word person name	0.06%

Comparative experiments were also performed with and without the removal of negative examples, in order to prove the increase in performance after applying random undersampling to the data. With random undersampling, random instances of the majority class are removed from the dataset in order for their number to reach that of the positive classes. The classification results, after applying the undersampling procedure, are presented in the second column of table 5.

Testing of the algorithm was performed using 10-fold cross validation.

Table 4. Comparative results for different context window sizes.

Class	F-score (-1,+1)	F-score (-2,+2)
NULL	0.969	0.96
AE	0.728	0.683
ME	0.557	0.64
TE	0.768	0.74
AP	0.851	0.767
MP	0.865	0.852
TP	0.84	0.774
E	0.667	0.621
AAM	0.754	0.675
MAM	0.769	0.708
TAM	0.611	0.643
AO	0.353	0.465
MO	0.194	0.293
TO	0.143	0.35
AT	0.911	0.802
MT	0.588	0.857
TT	0.939	0.818
AX	0.585	0.558
TX	0.588	0.492
AL	0.421	0.449
ML	0.059	0.17
TL	0.278	0.293
X	0.452	0.457
F	0.889	0.947
AN	0.286	0.364
TN	0.378	0.632
MX	0.524	0.561
MN	0	0
ON	0	0
N	0.667	0.571
L	0.519	0.506

Table 5. Column A: Results with two-phase learning for context window size (-1,+1). Column B: Results with two-phase learning for context window size (-1,+1) after applying random undersampling.

Class	F-score A	F-score B
NULL	0.981	0.939
AAM	0.895	0.895
AE	0.882	0.899
AL	0.651	0.571
AN	0.65	0.756
AO	0.81	0.85
AP	0.96	0.96
AT	0.985	0.98
AX	0.755	0.806
E	0.721	0.803
F	0.944	1
L	0.55	0.565
MAM	0.944	0.911
ME	0.831	0.808
ML	0.562	0.632
MN	0	0
MO	0.55	0.5
MP	0.957	0.963
MT	0.952	0.952
MX	0.802	0.8
N	0.533	0.571
ON	0	0
TAM	0.865	0.838
TE	0.871	0.903
TL	0.524	0.465
TN	0.65	0.579
TO	0.629	0.611
TP	0.932	0.932
TT	0.954	0.96
TX	0.736	0.774
X	0.567	0.694

5 Discussion

The context window size plays an important role in classification performance. Certain types of entities require a larger window for their accurate detection, while larger context is misleading for other types. To the former category belong the more ‘straightforward’ types, that are either normally introduced by one characteristic word or acronym/abbreviation, like person names and company names, or that end in one specific symbol or acronym/abbreviation, like monetary amounts and percentages.

As can be seen in table 4, classification for certain types reaches a poor score. Looking more closely at table 3, this can be attributed without a doubt to the sparseness that characterizes these types (multi-word person names, multi-word stock/bond names, multi-word locations). An interesting exception to this rule is newspaper/journal names, that reach very high scores, despite their low frequency, because they are normally introduced by specific words like 'εφημερίδα' (newspaper) or 'περιοδικό' (journal).

Table 4 also shows the high f-score achieved for the negative (NULL) class compared to that of the positive classes, due to its high over-representation in the dataset.

The first column of table 5 shows the positive effects of stacking on the task at hand. The f-score increases up to more than 50% after applying two-phase learning. This improvement is due to two reasons: First, the sequential nature of the class label tags (start, middle, end). The class of one entity depends largely on the class of the preceding and the following entities. Second, the inclusion of the predicted class of the candidate entity (from the previous learning stage) in the feature vector of the second stage forces the classifier to focus on the mistakes it made, and try to correct them. Difficult cases like multi-word locations and multi-word names are now dealt with satisfactorily.

Random undersampling also proved highly beneficial for the majority of the entity categories. It forces the learner to pay more attention to the minority classes. The random nature of the undersampling process is the reason that the results for certain entity types were not improved, as certain useful negative examples may have been removed.

One-word stock/bond names (ON) occur extremely seldom in the corpus. Person names consisting of more than two words (MN), are even more rare. The learner has not been able to detect these classes due to the sparseness.

Given, however, the nature and complexity of the corpus, the low level of preprocessing (compared to previous approaches that use phrase-chunked input), and the large number of class labels, the results of table 5 are very impressive when compared to the ones reported in the literature (section 1).

6 Conclusion

This paper has presented a set of methodologies that were applied to a Modern Greek economic corpus in order to help detect and label semantic entities in the economic domain that are important for information retrieval, data mining, question-answering systems. Unlike previous approaches to named-entity recognition, the present work aims at identifying a wider range of entities (apart from names of persons, organizations and locations), that are linked to the economic domain, like names of stocks, of newspapers, of banks, quantities, percentages, etc. Stacking was performed to help the instance-based classifier to focus on the tricky cases and learn from previous mistakes, leading thereby to a significant increase in accuracy. Another novel feature of the present work is the way it deals with the imbalance in the class distribution in the dataset. Further performance improvement was achieved after balancing the class distribution using undersampling of the majority class instances. The above techniques deal very well with the large number of class labels, with the low level of preprocessing, as well as the complicated nature of the corpus.

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.

References

1. Ciaramita, M., Altun, Y.: Named Entity Recognition in Novel Domains with External Lexical Knowledge. In Workshop on Advances in Structured Learning for Text and Speech Processing (NIPS) (2005)
2. Daelemans, W., van den Bosch, A., Zavrel, J.: Forgetting Exceptions is Harmful in Language Learning. Machine Learning, Vol. 34, (1999) 11-41
3. Hendrickx, I., van den Bosch, A.: Memory-based One-step Named-entity Recognition: Effects of Seed List Features, Classifier Stacking and Unannotated Data. Proceedings of the 7th Conference on Computational Natural Language Learning (CoNNL), Edmonton, Canada (2003)
4. Kermanidis, K., Fakotakis, N., Kokkinakis, G.: DELOS: An Automatically Tagged Economic Corpus for Modern Greek. In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC), Las Palmas de Gran Canaria (2002) 93-100
5. Kubat, M., Matwin, S.: Addressing the Curse of Imbalanced Training Sets. Proceedings of the International Conference on Machine Learning (ICML) (1997) 179- 186.
6. Laurikkala, J.: Improving Identification of Difficult Small Classes by Balancing Class Distribution. Proceedings of the 8th Conference on Artificial Intelligence in Medicine in Europe. Cascais, Portugal (2001) 63-66
7. Radu, F., Ittycheriah A., Jing H., Zhang T.: Named Entity Recognition through Classifier Combination. Proceedings of the 7th Conference on Computational Natural Language Learning (CoNNL), Edmonton, Canada (2003) 168-171
8. Sgarbas, K., Fakotakis, N., Kokkinakis, G.: A Straightforward Approach to Morphological Analysis and Synthesis, In Proceedings of the Workshop on Computational Lexicography and Multimedia Dictionaries (COMLEX), Kato Achaia, Greece (2000) 31−34
9. Sporleder, C., van Erp, M., Porcelijn, T., van den Bosch, A., Arntzen, P.: Identifying Named Entities in Text Databases from the Natural History Domain. In Proceedings of the 5th International Conference on Language Resources and Evaluation (2006)
10. Tsukamoto, K., Mitsuishi, Y., Sassano, M.: Learning with Multiple Stacking for Named Entity Recognition. In Proceedings of the 6th Conference on Natural Language Learning, Taipei, Taiwan (2002) 1-4
11. Wu. C., Jan, S., Tsai, T., Hsu, W.: On Using Ensemble Methods for Chinese Named Entity Recognition. Proceedings of the 5th SIGHAN Workshop on Chinese Language Processing, Sydney, Australia (2006) 142-145