Learning Shallow Syntactic Dependencies from Imbalanced Datasets: A Case Study in Modern Greek and English

Argiro Karozou, Katia Lida Kermanidis

Department of Informatics, Ionian University
7 Pl. Tsirigoti, 49100 Corfu, Greece
argykaroz@gmail.com, kerman@ionio.gr

Abstract. The present work aims to create a shallow parser for Modern Greek subject/object detection, using machine learning techniques. The parser relies on limited resources. Experiments with equivalent input and the same learning techniques were conducted for English, as well, proving that the methodology can be adjusted to deal with other languages with only minor modifications. For the first time, the class imbalance problem concerning Modern Greek syntactically annotated data is successfully addressed.

Keywords: shallow parsing, Modern Greek, machine learning, class imbalance

1 Introduction

Syntactic analysis is categorized into full/deep parsing -where a grammar and a search strategy assign a complete syntactic structure to sentences- and shallow parsing -finding basic syntactic relationships between sentence elements [15]. Information extraction, machine translation, question-answering and natural language generation are widely known applications that require shallow parsing as a pre-processing phase.

Shallow parsing may be rule-based [1][12][2] or stochastic [8][5]. Rule-based approaches are expensive and labor-intensive. Shallow parsers usually employ techniques originating within the machine learning (or statistical) community [19]. Memory Based Sequence Learning (MBSL) [3][18], and Memory-based learning (MBL) [9][21] have been proposed for the assignment of subject-verb and object-verb relations. The authors in [9] (the approach closest to the one described herein) provide an empirical evaluation of the MBL approach to syntactic analysis on a number of shallow parsing tasks, using the WSJ Treebank corpus [17]. Their reported f-measure is 77.1% for subject detection and 79.0% for object detection. The same techniques have been implemented for Dutch [4]. Regarding Modern Greek (MG), there is meager work in parsing and most of it refers to full parsing. Chunking and tagging have been attempted using Transformation-based error-driven learning [20].

This led to the idea of creating a shallow parser for MG. The present work deals with finding subject-verb and object-verb syntactic relations in MG text. A unique label (tag) is assigned to each NP-VP pair in a sentence. In an attempt to research the language-independence of the methodology, it is applied to English text as well, and

its performance in the two languages is compared. Furthermore, a basic problem is addressed, namely the imbalance of the learning examples of each class in the data, the so-called *class imbalance problem*. State-of-the-art techniques, like resampling, are employed for the first time to the authors' knowledge, to deal with this data disproportion in the task at hand, and the results are more than encouraging. Finally, the approach relies on limited resources, i.e. elementary morphological annotation and a chunker that uses two small keyword and suffix lexica to detect non-overlapping phrase chunks. Thereby the methodology is easily adaptable to other languages that are not adequately equipped with sophisticated resources.

MG has a rich morphology and does not follow the subject-verb-object (SVO) ordering schema. For instance, /ipia gala xthes/ (I drank milk yesterday), /gala ipia xthes/ and /xthes ipia gala/ are all syntactically correct and semantically identical. MG is a pro-drop (pronoun drop) language since the subject may be omitted. In MG verbs agree with their subject in gender and number. Subjects and predicates (nominals denoting a property of the subject, also called "copula") are in the nominative case, objects in the accusative and genitive case.

2 Data Collection

The MG text corpus used for the experimental process comes from the Greek daily newspaper "Eleftherotypia" (http://www.elda.fr/catalogue/en/text/W0022.html) and includes 3M words. A subset of the corpus (250K words) is morphologically annotated and automatically chunked by the chunker described in [20].The present methodology focuses on the identification of syntactic relations concerning the subject and object relations that NPs have with VPs in the same sentence. This restriction possibly excludes useful relations, the central meaning of a sentence, however, can be retrieved quite well considering just these relations.

Each NP-VP pair in a corpus sentence constitutes a learning instance. During feature extraction, i.e. the morphosyntactic features (21 in number) that represent the pair and its context and affect subject and object dependencies between an NP and a VP were selected: the case, the number and the person of the headword of the NP, the person, the number and the voice of the head word of the VP, the distance (number of intervening phrases) between the NP and VP, the verb type, i.e. whether it is connective (είμαι-to be, γίνομαι–to become, φαίνομαι–to seem), or impersonal (πρέπει-must, μπορεί-may, βρέχει-to rain, πρόκειται–about to be, etc.), the part of speech of the headword of the NP and VP, the number of commas, verbs, coordinating conjunctions and other conjunctions between the NP and VP, and the types of the phrases up to two positions before and after the NP and the VP. The dataset consisted of 20,494 learning instances, which were manually annotated with the correct class label ("Subject" for a subject-verb dependency, "Object" for a verb-object dependency, and "NULL" otherwise). The annotation process required roughly about 50 man-hours.

The same process was applied to an English corpus so as to compare and contrast it to the Greek shallow parser, and find out how well a parser, that utilizes equivalent resources and the same methodology can work for English. The corpus used was

Susanne (www.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/nlp/corpora/susanne), about 130K words of written American English text. Syntactic tree structures were flattened into IOB-format sentence structures of consecutive non-overlapping chunks, using Buchholz's software (http://ilk.uvt.nl/team/sabine). Thereby, the input data is equivalent for both languages. 6,119 NP-VP pairs were transformed into learning instances, and were manually annotated with one of the three class labels. Features were selected to describe morphosyntactic information about the NP-VP pair, taking into account the relevant properties of the language: the voice of the VP, whether the VP is infinitive or has a gerund form, its number and person, the part of speech of the headword of the NP and VP, the number of the headword of the NP, the distance (number of intervening phrases) and the number of conjunctions and verbs between the NP and VP and the types of the phrases up to two positions before and after the NP and the VP.

A disproportion of the class distribution in both datasets was clearly noticeable. In the MG data, from the 20,494 total learning instances, 16,335 were classified as null and only 4,150 were classified as subject or object while in the English data, from the 6,119, 4,562 were classified as null and just 1,557 as subject or object. This problem, where one or more classes are under-represented in the data compared to other classes, is widely known as Class Imbalance.

Class imbalance is a challenge to machine learning and data mining, and is prevalent in many applications like risk management, fraud/intrusion detection, text classification, medical diagnosis/monitoring, etc. Numerous approaches have been proposed both at the data and algorithmic levels. Concerning the data level, solutions include different kinds of re-sampling [16] [22]. In under-sampling, a set of majority instances is removed from the initial dataset while all the minority instances are preserved. In over-sampling, the number of minority instances is increased, so that they reach the number of majority instances. Random over-sampling in general is among the most popular sampling techniques and provides competitive results.

In this work, we employed random over-sampling by duplicating minority examples, random under-sampling, and feature selection [23], using filtering techniques available in the Weka machine learning workbench [24]: Resample and Synthetic Minority Over-sampling Technique (SMOTE) [13]. For high-dimensional data sets, filters are used that score each feature independently based on a rule. Resample produces a random subsample of a dataset, namely sampling with replacement. SMOTE is an over-sampling method [6] that generates synthetic examples in a less application-specific manner, by operating in "feature space rather than data space". Our final results show that feature selection is a very competitive method, as proven by different approaches [7].

3 Experimental Setup, Results and Discussion

The experiments were conducted using WEKA. After experimentation with different algorithms, those leading to the best results were selected: C4.5 (decision tree learning) and k-*nearest-neighbor (*instance-based learning, with $1 \leq k \leq 17$). Grafting [11] is also applied as a post process to an inferred decision tree to reduce the

prediction error. Post-pruning was applied to the inducted decision trees. Classifier performance was evaluated using precision, recall, and the f-measure. Validation was performed using *10 – fold cross validation*. The best results were obtained by the decision tree classifiers, and in particular with grafting.

As regards k-NN, the value of k affected significantly the results, while k increased more training instances that were relevant to the test instance were involved. Only after a relatively large value of k (k=17), the f-measure started dropping and performance was affected by noise. Results are listed below.

Table 1. First experiment for the MG and English corpora (classifiers evaluated by f-measure)

	C4.5	C4.5graft	1-NN	3-NN	9-NN	15-NN	17-NN
Subject (MG)	68.4	68.7	59.2	60.8	62	62.5	60.7
Object (MG)	77.1	77.4	60.7	64	65.5	69	65.8
Subject (Eng)	73.3	73.5	62.1	63.8	64.7	66.6	66.1
Object (Eng)	68.5	68.4	51.2	56.71	61.8	61	61.1

After the implementation of the methods that face the class imbalance problem, results had a significantly upward course. Concerning both datasets, the best results were achieved with C4.5 and are presented in Table 2.

Regarding the pre-processing phase, the outcome results are influenced strongly by the automatic nature of the chunking process. A small but concrete number of errors are attributed to erroneous phrase splitting, excessive phrase cut-up and erroneous phrase type identification.

Table2. Overall comparative results for the MG and English corpora evaluated by f-measure

	First Approach	Resample	Undersampling	Oversampling	SMOTE
Object (MG)	77.1	97.2	95.9	89.9	86
Subject (MG)	68.4	94.2	86.4	87.9	68
Object (Eng)	68.5	93.6	79.4	85.8	81.4
Subject (Eng)	73.3	92.1	83	88.1	73.4

A significant difference between subject and object detection performance in the Greek corpus was also noticed. Object detection exceeded subject detection by almost 10%. This is due to the existence of noise. In many cases training instances had exactly the same feature description, but different classification. It is a general problem that pertains to the difficulty of the MG language and especially the problem of distinguishing the subject of a copular (linking) verb from its copula. The instances of the copula and the subject had the same features (nominative case, short distance from the verb, etc.) but different class values. In the MG sentence *NP [Το θέμα της συζήτησης] VP[είναι] NP[τα σκουπίδια] (NP[the point of the conversation] VP[is] NP[the rubbish]),* the first NP is the subject and the second the copula. The same syntactic relation would hold even if the NPs were in different order.

Classifiers behaved almost in the same way on the English corpus. This shows that the shallow parser can be adjusted (using a very similar set of features and equivalent pre-processing) to cope with the English language. Better results were obtained for the subject recognition in English. This indicates that subject detection is easier in

English, as it has a stricter and simpler structure. Compared to previous work [9], the results presented herein are very satisfying and outperform those of similar approaches (the authors in [9] report an accuracy of 77.1% for subject and 79% for object recognition).

In both languages, over-sampling led to the second-best results after the Resample method, except for the case of object detection in the MG corpus, meaning that during random under-sampling more noisy instances were removed. Generally, as over-sampling gave better results than under-sampling, it turns out that under-sampling is not always as effective as has been claimed [6].

During the experimental process for the MG data, a lexicalized version of the data was also used, according to the approach of Daelemans et al. [9]. The lemma of the head word of the VP was included in the data. However, the final results dropped significantly (the f-measure for C4.5 was 61.1% for subject and 68.3% for object detection). The verb type used in our initial feature set is sufficient for satisfactory classification, while lexicalization includes redundant information that is misleading.

Important future improvements could be: the application of other techniques to deal with the class imbalance problem (e.g. cost-sensitive classification, focused one-sided sampling etc.), an improvement of the second level of the shallow parser, experimentation with a larger number of training instances. The creation of a fourth class for the classification of NPs that are copulas could improve the final result if the feature set was altered accordingly, and force classifiers to learn to distinguish between these two types relations (subjects and copulas) and finally the recognition of other types of syntactic relations of the verb and simultaneously the inclusion of other phrase types, apart from NPs.

4 Conclusion

In this paper, a shallow parser for MG subject/object detection is created, using machine learning algorithms. The parser utilizes minimal resources, and does not require grammars or lexica of any kind. The parser can be adjusted to cope with English language text as well, with minor modifications, with impressive results. Additionally, for the first time to the authors' knowledge, the class imbalance problem in the data is successfully addressed and the final results climb up to 97.2% for object and 94.2% for subject detection in the MG corpus and 92.1% and 93.6% respectively in the English corpus.

References

1. Abney S: In Principle-Based Parsing: Computation and Psycholinguistics, pp. 257-278. Kluwer Academic Publishers, (1991)
2. Aït-Mokhtar S., Chanod P.: Subject and Object Dependency Extraction Using Finite-State Transducers. Rank Xerox Research Centre, France
3. Argamon, D.I. and Krymolowski Y.: A Memory-based Approach to Learning Shallow Natural Language Patterns. 36th Annual Meeting of the ACL, pp. 67-73, Montreal (1998)

4. Canisius S.: Memory-Based Shallow Parsing of Spoken Dutch. MSc Thesis. Maastricht University, The Netherlands (2004)
5. Charniak.E.: Statistical Parsing with a Context-free Grammar and Word Statistics. In Proc. National Conference on Artificial Intelligence (1997)
6. Chawla N., Japkowicz N., Kolcz A.: Special Issue on Learning from Imbalanced Data Sets. Sigkdd Explorations, Canada (2005)
7. Chen Y.: Learning Classifiers from Imbalanced, Only Positive and Unlabeled Data Sets. Department of Computer Science Iowa State University (2009)
8. Collins M.: Three Gneretive, Lexicalised Models for Statistical Parsing. Univerity of Pennsylvania, U.S.A (1996)
9. Daelemans W., Buchholz S., Veenstra J.: Memory-based Shallow Parsing, ILK, Tilburg University (2000)
10. Evaluations and Language resources Distribution Agency, http://www.elda.fr/catalogue/en/text/W0022.html
11. Geoffrey I. Webb.: Decision Tree Grafting From the All-Tests-But-One Partition. Deakin University, Australia
12. Grefenstette G.: Light parsing as finite-state filtering. In Wolfgang Wahlster, editor, Workshop on Extended Finite State Models of Language. ECAI'96, Budapest, Hungary. John Wiley & Sons, Ltd. (1996)
13. Hulse J., Khoshgoftaar T., Napolitano A.: Experimental Perspectives on Learning from Imbalanced Data. Florida Atlantic University, Boca Raton, USA (2007)
14. Journal of Machine Learning Research http://jmlr.csail.mit.edu/papers/special/shallow_parsing02.html
15. Jurafsky D., Martin J.: Speech and Language Processing: An Introduction to Natural Processing, Computational Linguistics, and Speech Recognition (2000)
16. Ling C. X. and Li C.: Data Mining for Direct Marketing: Problems and Solutions. American Association for Artificial Intelligence. Western Ontario University (1998)
17. Marcus et al.: Building a large annotated corpus of English: The penn Treebank. Coputational Linguistics, 19(2): 313-330 (1993)
18. Munoz M. et al.: A Learning Approach to Shallow Parsing, Department of Computer Science University of Illinois at Urbana (1999)
19. Roth D. and Yih W.: Probabilistic reasoning for entity & relation recognition. In Proc. Of COLING-2002, pages 835–841 (2002)
20. Stamatatos, E., Fakotakis, N. and Kokkinakis, G.: A Practical Chunker for Unrestricted Text, Proceedings of the Conference on Natural Language Processing, Patras, Greece, pp. 139-150 (2000)
21. Sang T. K, E. F. and Veenstra, J.: Representing text chunks. In Proceedings of EACL 1999, pp. 173–179 (1999)
22. Solberg, A., & Solberg, R.: A Large-Scale Evaluation of Features for Automatic Detection of Oil Spills in ERS SAR Images. In International Geoscience and Remote Sensing Symposium, pp. 1484-1486 Lincoln, NE (1996)
23. Wasikowski M.: Combating the Small Sample Class Imbalance Problem Using Feature Selection. In: 10th IEEE Transactions on Knowledge and data engineering (2010)
24. Witten I. & Eibe F.: Data Mining: Practical Machine Learning Tools and Techniques. Department of Computer Science University of Waikato (2005)