Learning to Case-Tag Modern Greek Text

Antonis Koursoumis, Evangelia Gkatzou, Antigoni M. Founta, Vassiliki I. Mavriki, Karolos Talvis, Spyros Mprilis, Ahmad A. Aliwat, and Katia Lida Kermanidis

Ionian University, Department of Informatics
7 Tsirigoti Square, 49100, Corfu, Greece
kerman@ionio.gr

Abstract. Morphological case tagging is essential for the identification of the syntactic and semantic roles of sentence constituents in most inflectional languages. Although it is usually viewed as a side-task of general tagging applications, it is addressed in the present work as an individual, stand-alone application. Supervised learning is applied to Modern Greek textual data in order to case-tag declinable words using merely elementary lexical information and local context. Several experiments with various context window sizes, as well as base- and meta-learning schemata, were run with promising results.

Keywords: tagging, morphological case, supervised learning, ensemble learning, Modern Greek.

1 Introduction

Tagging is the process of assigning morphological attributes to words in natural language text. Tagged text can then be utilized for further analysis, i.e. parsing, information extraction etc. A significant part of previous approaches to automatic tagging have focused on the identification of the part-of-speech (pos) tag of a word using rule-based systems, statistical models, supervised and unsupervised learning algorithms, and combinations of the above. Another set of approaches addresses broader morphological information, where the tag set includes gender, number, case, person, voice, word type and other elements of grammatical information.

Only very limited approaches have given special attention to or, much less, have focused solely on case tagging. These approaches addressed morphologically rich, inflectional languages, where words in certain pos categories (usually nominals) may appear in several cases. The most usual grammatical cases, that are present in the vast majority of inflectional languages, are the nominative, the genitive and the accusative. However, certain languages, contemporary or ancient, have special cases, like the dative, the vocative, the ablative, the instrumental, the locative.

The grammatical case of a nominal element is quite significant as it determines the relationship between the nominal and its head (e.g. a noun with the main verb, a modifier with the noun it modifies etc.), and it is therefore essential for identifying the syntactic and semantic roles of the elements in a sentence.

Unlike previous approaches, the present work focuses solely on case tagging for the first time in Modern Greek (MG). The methodology makes use of low-level

I. Maglogiannis, V. Plagianakos, and I. Vlahavas (Eds.): SETN 2012, LNAI 7297, pp. 353–360, 2012.

information, i.e. elementary morphological information. No lexica, lemmatizers, stemmers, or stress identification tools are made use of. The context is taken into account, without exploiting any kind of information regarding its syntactic structure. Thereby, the methodology is not resource-demanding and may easily be adapted to other languages that have a case-based morphology. Supervised learning has been employed for the prediction of the correct grammatical case value. A novel feature-vector structure has been proposed for the learning instances and a set of state-of-the-art learning algorithms, stand-alone- as well as meta-classifiers, have been experimented with, and their performance is compared.

The rest of this paper is organized as follows. Section 2 describes previous approaches related to case tagging. Section 3 introduces some important MG properties that affect the task. The dataset, the experimental setup and the results are presented and discussed in section 4. The paper concludes in the last section.

2 Related Work

As mentioned before, there are significant ambiguity problems during the process of tagging because of the idiosyncrasies of some languages. This kind of problem does not seem to appear only in MG, but in many morphologically rich languages like Arabic, Turkish, Dutch and Icelandic. Several efforts have been made in the last decade to address this problem.

In Icelandic, a case tagging approach has been proposed [2], who developed a case tagger for non-local case and gender decisions, using 639 tags and approximately a corpus of 590.000 tokens. The approach achieved an accuracy of 97.48%.

In the case of Arabic, a memory-based learning for morphological analysis and pos tagging has been proposed, using as input unvoweled words [8]. The tagger reaches an accuracy rate of 93.3% for known and 66.4% for unknown words.

A similar approach has been proposed for a morphosyntactic tagger and dependency parser for Dutch called TADPOLE [16]. TADPOLE is 96.5% correct on known and 79.0% on unseen words.

When it comes to the Greek language, many difficulties are yet to be surpassed, although certain important approaches have already been published [9][10][11][12], focusing almost exclusively on pos tagging.

Finally, for the Turkish language some approaches have been proposed in the past years. The most important attempt implemented a system that extracts a corpus from the Web, annotates its sentences with case information and uses the Naïve Bayes classifier to convey subcategorization frames (SFs) on Turkish verbs [15] with promising results, taking into account the properties of the language.

3 Morphological Case in Modern Greek

MG has a complex inflectional system. The morphological richness allows for a relatively free-word-order syntax, where the role of each constituent is determined by its morphological features rather than by its position in the sentence.

There are eleven different pos categories in MG, six declinable (articles, nouns, adjectives, pronouns, verbs and numerals), and five indeclinable (adverbs, prepositions,

conjunctions, interjections and particles). All indeclinable words plus articles and pronouns form closed sets of words, while nouns, adjectives, and verbs form open sets. Nouns, adjectives, pronouns, numerals and participles are characterized by the case attribute. Its possible values are: nominative, genitive, accusative and vocative. The dative case was extensively used in Ancient Greek, but its function has been taken over either by other cases or by other syntactic structures (e.g. certain prepositional phrases) in MG ("Syncretism").

Morphological cases indicate types of syntactic relations. A nominal element in the nominative case denotes a subject, or the copula of a linking verb, the accusative case denotes a direct object or a temporal expression, while the genitive case denotes possession or an indirect object. In prepositional phrases, the case of the nominal element (genitive or accusative) depends on the preposition introducing the phrase.

The morphological richness renders pos ambiguity a less significant research challenge than case ambiguity in MG. The same word form may often be found in text associated with different sets of morphological features. In other words, the same word form can appear in texts having three different case values. Very often declinable words have the same orthographic form in the nominative and in the accusative case. This holds for almost all nouns, adjectives, articles, pronouns and ordinal numerals, feminine and neutral, singular and plural. For example, the phrase

	Ακούει	το	παιδί
(S(he))	listens	the	child

has two different meanings: "The child is listening" and "Someone is listening to the child". The first meaning sets the noun phrase "το παιδί" (the child) in the nominative case (the child as a subject), while, the second, in the accusative (the child as the object). Another common ambiguity is the genitive and the accusative case. This holds for most singular masculine and neutral nouns and adjectives. Also, the vocative case shares the same orthographic form with the genitive and the accusative and the genitive case in several masculine nouns and adjectives, and with the nominative and accusative case in several feminine and neutral nouns and adjectives.

4 Morphological Case Learning

The MG corpus used for the experiments comes from the Greek daily newspaper "Eleftherotypia" [3]. The subset of the corpus (250K words) used for the experiments described herein is manually annotated with morphological information.

The dataset we use for our calculations and experiments consists of 65535 instances. Each instance corresponds to a specific word in the corpus that has a case; it holds info referring to the two words following and the two words preceding the focus word. The dataset uses thirty four different attributes, one of which is the class of the focus word. The attributes are listed in Table 1. The class of the focus word takes three different values: "n", "g", "a", "v" and "d", corresponding to the nominative, genitive, accusative, vocative and dative case respectively. The analysis of the dataset shows that 30,7% of the words are in the nominative, 26,2% in the genitive and 42,9% in the accusative case. The vocative and dative instances are extremely rare, i.e. 0,04% and 0,06% respectively.

Table 1. The features of the learning vector

	Feature Description		Feature Description
1	last three letters of the focus word	18	last three letters of word+1
2	last two letters of the focus word	19	last two letters of word+1
3	focus word pos	20	word+1 pos
4	focus word gender	21	word+1gender
5	focus word number	22	word+1 number
6	last three letters of word-1	23	last three letters of word+2
7	last two letters word-1	24	last two letters of word+2
8	word-1pos	25	word+2 pos
9	word-1 gender	26	word+2 gender
10	word-1 number	27	word+2 number
11	word-1 case (manual)	28	position of the closest previous verb
12	last three letters of word-2	29	number of the closest previous verb
13	last two letters of word-2	30	voice of the closest previous verb
14	word-2 pos	31	position of the closest next verb
15	word-2 gender	32	number of the closest next verb
16	word-2 number	33	voice of the closest next verb
17	word-2 case (manual)	34	focus word case (class)

State-of-the-art learning algorithms have been experimented with. The Weka workbench (http://www.cs.waikato.ac.nz/ml/weka) was used for the experiments presented herein. Tests were run with instance-based learning (k-NN) in order to determine the best value for k. Decision trees (C4.5 for unpruned trees as well as with reduced error pruning) were created; the Naive Bayes classifier was also tested. Apart from these stand-alone classifiers, ensemble learning schemata have also been experimented with. Stacking, boosting and bagging were run on the data with promising results. For stacking, the prediction results of three stand-alone classifiers (Naïve Bayes, 7-NN and C4.5) are taken as input to the meta-learner (C4.5). For bagging, C4.5 was applied iteratively ten times, each time to a randomly chosen portion (60%) of the data. For boosting, AdaBoost was run with C4.5 and reduced-error pruning. The context window size has been experimented with as well: four different datasets were created, i.e. (-2,+2), (-1,+1), (-2,0), (-1,0). Validation was performed using 10-fold cross validation. Classification results (precision and recall for all class values, classifiers and context window sizes) are shown in Figure 1.

As is to be expected, results for the vocative and dative cases are poor due to their sparse occurrence in the data. The dative case scores somewhat better and in some cases (e.g. with C4.5) quite well, because of the characteristic orthographic endings of the words in this case (usually archaic expressions from Ancient Greek or other earlier versions of the Greek language, that are still used). The advanced meta-learning schemata are also able to capture these orthographic idiosyncrasies quite well despite its sparseness. The majority of vocative examples are classified as nominative, due to the large ambiguity problem, as described in section 3, on top of the sparseness. Regarding the remaining cases, the genitive achieves the highest results, i.e. 98% with boosting. The genitive is rarely mistaken as one of the other cases; the ambiguity is slightly higher between the nominative and the accusative.

Regarding the context window, preceding words are more important than the ones following the focus word, probably due to the preceding articles, modifiers and determiners that often share their case with the following headword. The following words, and especially word+2, seem to be more misleading than helpful. According to the

information gain ratio, the most important features for learning are the ending letters and pos of the focus word as well as the morphological features of the previous words (especially the word-1 case), the ending letters of word-1 and the voice of the closest verb.

Concerning the algorithms, the larger the learning vector, the more neighbors are required for accurate instance-based learning. Using the two preceding words, and a large number of neighbors, k-NN is able to learn the sparse cases with flawless precision. Pruned C4.5 performs significantly better than without pruning, especially for the sparse cases. Naïve Bayes leads to the poorest overall results, while the metalearners (especially boosting) lead to the best results regarding the vocative case.

Placing the current approach on the map of related MG text tagging approaches [9][10][11][12], the results reported herein are comparable to the ones reported in the

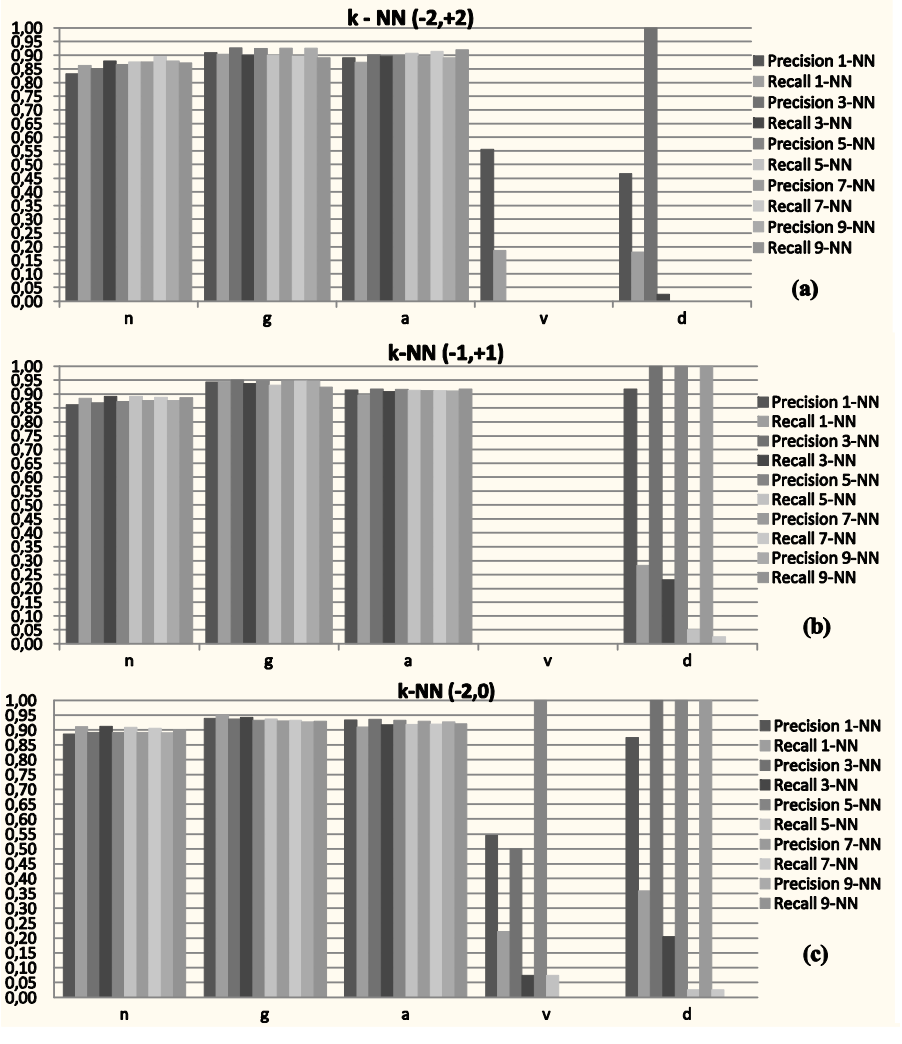

Fig. 1. (a)-(j). Experimental results

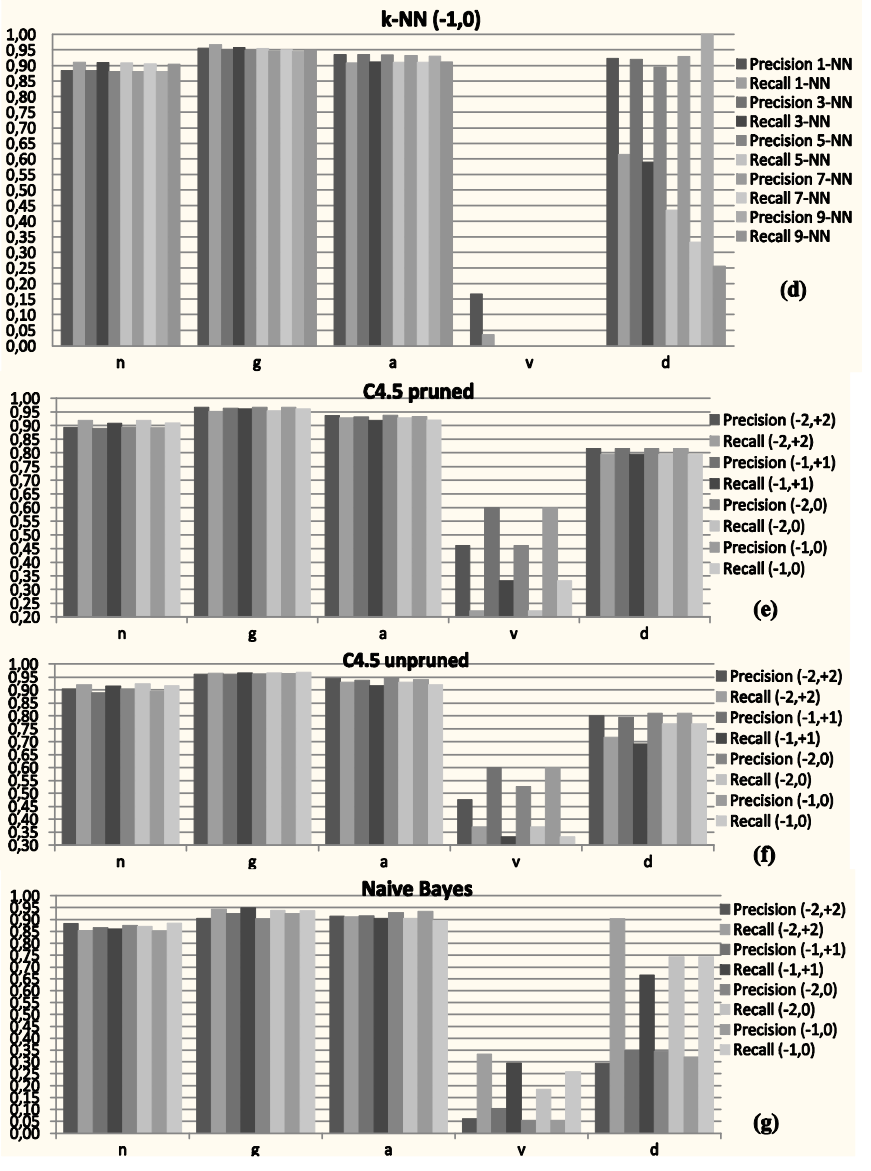

Fig. 1. (*continued*)

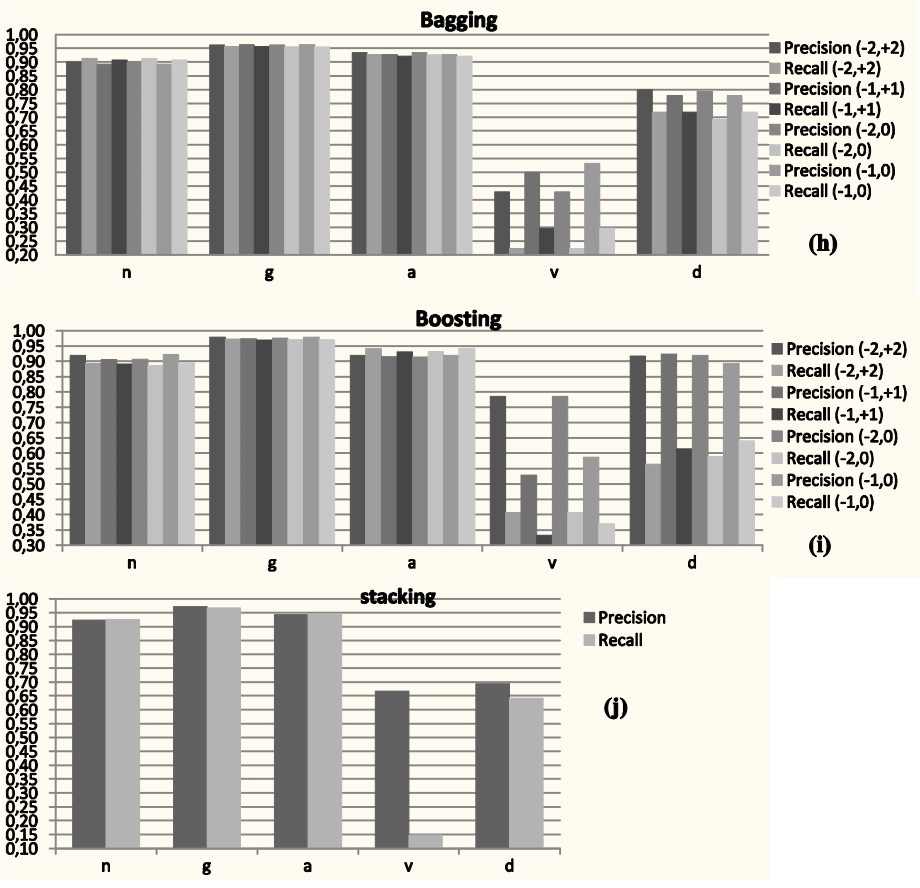

Fig. 1. (*continued*)

literature, even though no direct comparison is feasible, as most previous work is on pos tagging. Their reported accuracy varies from 85% to 98.2%, the corpus sizes vary from 140K to 1.9M words (much larger than ours), and employed resources include stemming (prefix and suffix identification), stress information, and/or morphological lexica, none of which is available in the present approach.

5 Conclusion

A methodology exclusively for morphological case tagging of MG has been presented. It relies on minimal resources, i.e. morphological and context information, and addresses satisfactorily the free word order of MG, as well as its rich inflectional system. Several learning algorithms and the context window surrounding the focus word have been experimented with, and the features that are significant for case learning have been investigated. Special morphosyntactic features that might help learning the sparse cases, feature selection pre-processing, higher-level resources (e.g. lemmatization, syntactic structures) would be interesting future research aspects to explore.

References

1. Brill, E.: Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. Computational Linguistics 21(24) (1995)
2. Dredze, M., Wallenberg, J.: Icelandic Data Driven Part of Speech Tagging. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Columbus, OH, USA, Short Papers 33–36 (2008)
3. Evaluations and Language Resources Distribution Agency, `http://www.elda.fr/catalogue/en/text/W0022.html`
4. Freeman, A.: Brill's POS tagger and a morphology parser for Arabic. In: ACL/EACL-2001 Workshop on Arabic Language Processing: Status and Prospects, Toulouse, France (2001)
5. KEME (Center of Educational Studies and Training in Education): Revision of Modern Greek Grammar of Manolis Triantafillidis (in Greek). Didactic Books Publishing Organization (1983)
6. Kramarczyk, I.: Improving the tagging accuracy of Icelandic text. MSc Thesis, Reykjavík University (2009)
7. Loftsson, H.: Tagging Icelandic Text using a Linguistic and a Statistical Tagger. In: NAACL-Short 2007 Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume, Short Papers 105–108 (2007)
8. Marsi, E., Van den Bosch, A., Soudi, A.: Memory-based morphological analysis and part-of-speech tagging of Arabic. In: Soudi, A., van den Bosch, A., Neumann, G. (eds.) Arabic Computational Morphology Knowledge-based and Empirical Methods. Springer (2007)
9. Orphanos, G., Tsalidis, C.: Combining Handcrafted and Corpus-Acquired Lexical Knowledge into a Morphosyntactic Tagger. In: Proceedings of the 2nd CLUK Research Colloquium, Essex, UK (1999)
10. Papageorgiou, H., Prokopidis, P., Giouli, V., Piperidis, S.: A Unified POS Tagging Architecture and its Application to Greek. In: Proceedings of Second International Conference on Language Resources and Evaluation, LREC 2000, Athens, Greece, pp. 1455–1462 (2000)
11. Papakitsos, E., Grigoriadou, M., Ralli, A.: Lazy Tagging with Functional Decomposition And Matrix Lexica: An Implementation in Modern Greek. Literary and Linguistic Computing 13(4), 187–194 (1998)
12. Petasis, G., Paliouras, G., Karkaletsis, V., Spyropoulos, C., Androutsopoulos, I.: Resolving Part-of-Speech Ambiguity in the Greek Language Using Learning Techniques. In: Proc. of the ECCAI Advanced Course on Artificial Intelligence, Chania, Greece (1999)
13. Shen, L., Satta, G., Joshi, A.K.: Guided learning for bidirectional sequence classification. In: ACL (2007)
14. Triantafillidis, M.: Modern Greek Grammar (Dimotiki) (in Greek). Reprint with corrections 1978. Institute of Modern Greek Studies, Thessaloniki (1941)
15. Usun, E., et al.: Web-based Acquisition of Subcategorization Frames for Turkish. In: 9th International Conference on Artificial Intelligence and Soft Computing. IEEE Computational Intelligence Society, Los Alamitos (2008)
16. Van den Bosch, A., Busser, G.J., Daelemans, W., Canisius, S.: An efficient memory-based morphosyntactic tagger and parser for Dutch. Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting, Leuven, Belgium, pp. 99–114 (2007)