Dealing with the grey sheep of the Romanian gender system, the neuter

Liviu P. DINU, Vlad NICULAE, Maria ŞULEA
University of Bucharest, Faculty of Mathematics and Computer Science,
Centre for Computational Linguistics, Bucharest
ldinu@fmi.unibuc.ro, vlad@vene.ro, mary.octavia@gmail.com

ABSTRACT
Romanian has been traditionally seen as bearing three lexical genders: masculine, feminine, and neuter, although it has always been known to have only two agreement patterns (for masculine and feminine). Previous machine learning classifiers which have attempted to discriminate Romanian nouns according to gender have taken as input only the singular form, either presupposing the traditional tripartite analysis, or using additional information from case inflected forms. We present here a tool based on two parallel support vector machines using n-gram features from the singular and from the plural, which distinguish the neuter.

KEYWORDS: Romanian morphology, neuter gender, noun class, SVM.

Proceedings of COLING 2012: Demonstration Papers, pages 119–124,
COLING 2012, Mumbai, December 2012.

1 The Romanian gender system

Recently, a big grammatical mistake made by a Romanian politician brought in attention the plural form of Romanian nouns. The traditional analysis (Graur et al., 1966; Rosetti, 1965, 1973; Corbett, 1991) identifies Romanian as the only Romance language bearing three lexical genders (masculine, feminine and neuter), whether the neuter was inherited from Latin (Constantinescu-Dobridor, 2001, p. 44), or redevelopped under the influence of Slavic languages (Rosetti, 1965; Petrucci, 1993). The first two genders generally have no problem regarding their plurals (follow a pattern more or less), the neuter gender being the one which poses some dificulties.These dificulties are not encompased only by politicians, but also for second language acquisition and; not to mention, in some cases, the long debates between linguists themselves. The problem occurs since the neuter gender has a masculine form for singular and a feminine form for plural (see Table 1 for examples). Since the language bears only two agreement markers (masculine and feminine), the three genders then need to be mapped onto the dual agreement, the way in which this mapping is done and on what basis also having been debated. However, under the premise that gender is expressed through agreement, the fact that Romanian neuter nouns lack their own marking and their own agreement pattern (they systematically and without exception follow the masculine agreement in the singular and the feminine in the plural as seen in Table 1) have lead Bateman and Polinsky (2010) and others to ask the question of whether Romanian has three genders, or two. Gender assignment thus becomes a burden not only for linguists to describe, but also for second language learners of Romanian to acquire.

	singular	plural
masculine	băiat **frumos** boy.M beautiful.M	băieți frumoși boy.M beautiful.M
neuter	creion **frumos** crayon.N beautiful.M	creioane **frumoase** crayon.N beautiful.F
feminine	fată frumoasă girl.F beautiful.F	fete **frumoase** girl.F beautiful.F

Table 1: Gender vs. agreement in Romanian

In our best knowledge, there are only two computational linguistics based approaches which attempted to discriminate Romanian nouns according to gender: Nastase and Popescu (2009) and (Cucerzan and Yarowsky, 2003). Our goal was, thus, to better -in comparison to Nastase and Popescu (2009)'s results- or successfully -in comparison to Cucerzan and Yarowsky (2003)'s experiment- distinguish these "neuter" nouns from feminines and masculines, by employing the minimum amount of information. We employed phonological information (coming from singular and plural noninflected nominative forms) as well as information coming from the feminine and masculine gender labels. In what follows we will present our tool for Romanian neuter nouns, which outperforms all previous attempts.

2 Our approach

We will look at singular and plural nominative indefinite forms (as specified by Bateman and Polinsky and used by Nastasescu and Popescu) and see if phonological features (endings) and information from masculine and feminine labels are sufficient to correctly classify Romanian neuter nouns as such. Another thing to take into consideration when looking at our classifier is the fact that, while Bateman and Polinsky (2010, p. 53-54) use both

semantic and phonological features to assign gender, with the semantic features overriding the formal, we were unable use any semantic features, and used their phonological form as training examples.

2.1 Dataset

The dataset we used is a Romanian language resource containing a total of 480,722 inflected forms of Romanian nouns and adjectives. It was extracted from the text form of the morphological dictionary RoMorphoDict (Barbu, 2008), which was also used by Nastase and Popescu (2009) for their Romanian classifier, where every entry has the following structure:

```
form␣lemma␣description
```

Here, 'form' denotes the inflected form and 'description', the morphosyntactic description, encoding part of speech, gender, number, and case. For the morphosyntactic description, the initial dataset uses the slash ('/') as a disjunct operator meaning that 'm/n' stands for 'masculine or neuter', while the dash ('-') is used for the conjunct operator, with 'm-n' meaning 'masculine and neuter'. In the following, we will see that some of the disjunct gender labels can cause some problems in the extraction of the appropriate gender and subsequently in the classifier. Since our interest was in gender, we discarded all the adjectives listed and we isolated the nominative/accusative indefinite (without the enclitic article) form. We then split them into singulars and plurals; the defective nouns were excluded. The entries which were labeled as masculine or feminine were used as training and validation data for our experiment, while the neuters were left as the unlabeled test set. The training and validation set contained 30,308 nouns, and the neuter test set 9,822 nouns (each with singular and plural form).

2.2 Classifier and features

Our model consists of two binary linear support vector classifiers (Dinu et al., 2012), one for the singular forms and another one for the plural forms. Each of these has a free parameter C that needs to be optimized to ensure good performance. We extracted character n-gram features vectors from the masculine and feminine nouns, separately. These vectors can represent counts of binary occurences of n-grams. We also considered that the suffix might carry more importance so we added the '$' character at the end of each inflected form. This allows the downstream classifier to assign a different weight to the $(n-1)$-grams that overlap with the suffix. Each possible combinations of parameters: n-gram length, use of binarization, addition of suffix, and the C regularization parameter was evaluated using 10-fold cross-validation, for both singular and plurals. After the model has been selected and trained in this manner, the neuter nouns are plugged in and their singular forms are classified according to the singular classifier, while their plural forms are classified by the plural model. The experiment was set up and run using the *scikit-learn* machine learning library for Python (Pedregosa et al., 2011). The implementation of linear support vector machines used is *liblinear*.

3 Our results

The best parameters chosen by cross-validation are 5-gram features, append the suffix character, but don't binarize the feature vectors. On masculine-feminine singulars, this

obtained an accuracy of 99.59%, with a precision of 99.63%, a recall of 99.80% and an F_1 score of 99.71%. The plural model scored an accuracy of 95.98%, with a precision of 97.32%, a recall of 97.05% and an F_1 score of 97.18%. We then moved on to check the classification results of the neuter forms, and performed error analysis on the results. Table 2a shows the distribution of neuter noun tuples (singular, plural) according to how our models classify their forms. Our hypothesis states that all of the mass should gather in the top-left corner, i.e. neuters should classify as masculine in the singular and feminine in the plural. There are more misclassifications in the plural form of neuter nouns than in their singular form. In what follows, we will briefly analyze the misclassifications and see if there is any room for improvement or any blatant mistakes that can be rectified.

s/p	f	m
m	8997	741
f	69	15

(a) With full training set

s/p	f	m
m	9537	201
f	83	1

(b) Misleading samples removed

Table 2: Distribution of neuters as classified by the system. In each table, the upper left corner shows nouns classified as expected (masculine in the singular, feminine in the plural), while the lower right corner shows completely misclassified nouns (nouns that seem to be feminine in the singular and masculine in the plural). The other two fields appropriately show nouns misclassified in only one of the forms.

3.1 Analyzing misclassifications

We first notice that 10 out of the 15 nouns that were completely misclassified are French borrowings which, although feminine in French, designate inanimate things. According to (Butiurca, 2005, p. 209), all feminine French nouns become neuter once they are borrowed into Romanian. The ones discussed here have the singular ending in 'é', written in Romanian without the accent, but retaining main stress as in French. Another of the 15, which also ends in an 'e' carrying main stress but not of French origin, is a noun formed from an acronym: *pefele* from *PFL*. There is also a noun (*coclaură–coclauri*) probably from the pre-Latin substratum, which is listed in Romanian dictionaries either as a pluralia tantum or as it is listed in the dataset. The others are feminine singular forms wrongly labeled in the original corpus as being neuter or neuter/feminine. Looking at the entries in the original dataset for two of the last five nouns completely misclassified (*levantin/levantină–levantinuri/levantine* and *bageac/bageacă–bageacuri/bageci*), we notice that the latter receives an 'n' tag for the singular form *bageacă*, which in (Collective, 2002) is listed as a feminine, and the former receives the 'n/f' tag, meaning *either a neuter, or a feminine* (Barbu, 2008, p. 1939), for both the neuter *levantin* and the feminine *levantină* singular form. We further notice that, when the gender tag 'n/f' accompanies a singular form, from the perspective of our system, a contradiction is stated. Seeing as Romanian has only two agreement patterns and that neuters agree like masculines in the singular and feminines in the plural, the feminine form levantină cannot be either neuter, and receive the masculine numeral *un* in the singular, or feminine, and receive the feminine numeral *o*. It can only be feminine. Through analoguous resoning, the tag 'n/m' accompanying a plural form is also "absurd". By eliminating the second gender from the two disjunct labels of the original dataset when extracting the nouns

for our classifier, we correctly tagged the neuter variants with 'n', but also wrongly tagged 5 feminine singular forms with 'n' and 7 masculine plural forms with 'n'. There are other misclassified nouns, from the other two groups, whose misclasification is due to an error in their initial gender label, for instance *algoritm–algoritmi* is shown to be a masculine in (Collective, 2002), however in the corpus it is tagged as neuter (together with the neuter variant *algoritm–algoritme*) and it subsequently appears to be misclassified in the plural as a masculine, which in fact it is. Another problem causing the misclassification is represented by the hyphenated compound nouns, which are headed by the leftmost noun that also receives the number/gender inflection. Seeing as our classification system weighed more on the suffix, it was prone to fail in correctly clasifying them.

Conclusion and perspectives

The results of our classifier make a strong case, in particular, for Bateman and Polinsky's analysis according to which class membership of nouns in Romanian is assigned based on form (nominative noninflected singular endings and plural markers), when semantic cues relating to natural gender (masculine and feminine) are absent, and, in general, for their two separate (for the singular and plural) dual-class division of the Romanian nominal domain. Furthermore, our classification model outperforms the two classifiers of Romanian nouns according to gender previously constructed in terms of correctly distinguishing the neuter.

Acknowledgments

The research of Liviu P. Dinu was supported by the CNCS, IDEI - PCE project 311/2011, "The Structure and Interpretation of the Romanian Nominal Phrase in Discourse Representation Theory: the Determiners." Note that the contribution of the authors to this paper is equal.

References

Barbu, A.-M. (2008). Romanian lexical databases: Inflected and syllabic forms dictionaries. In *Sixth International Language Resources and Evaluation (LREC'08)*.

Bateman, N. and Polinsky, M. (2010). *Romanian as a two-gender language*, chapter 3, pages 41–78. MIT Press, Cambridge, MA.

Butiurca, D. (2005). Influenţa franceză. In *European Integration-Between Tradition and Modernity (EITM), Volume 1*, pages 206–212.

Collective (2002). *Dicţionar ortografic al limbii române*. Editura Litera Internaţional.

Constantinescu-Dobridor, G. (2001). *Gramatica Limbii Române*. Editura Didactică şi Pedagogică Bucureşti.

Corbett, G. G. (1991). *Gender*. Cambridge University Press.

Cucerzan, S. and Yarowsky, D. (2003). Minimally supervised induction of grammatical gender. In *HLT-NAACL 2003*, pages 40–47.

Dinu, L. P., Niculae, V., and Şulea, O.-M. (2012). The romanian neuter examined through a two-gender n-gram classification system. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the*

Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey. European Language Resources Association (ELRA).

Graur, A., Avram, M., and Vasiliu, L. (1966). *Gramatica Limbii Române*, volume 1. Academy of the Socialist Republic of Romania, 2nd edition.

Nastase, V. and Popescu, M. (2009). What's in a name? in some languages, grammatical gender. In *EMNLP*, pages 1368–1377. ACL.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Petrucci, P. R. (1993). *Slavic features in the history of Romanian*. PhD thesis.

Rosetti, A. (1965). *Linguistica*. The Hague: Mouton.

Rosetti, A. (1973). *Breve Histoire de la Langue Rumain des Origines a Nos Jours*. The Hague: Mouton.