RoCo-News: A Hand Validated Journalistic Corpus of Romanian

Dan Tufiş, Elena Irimia

Romanian Academy Research Institute for Artificial Intelligence
13, Calea 13 Septembrie, Bucharest, 050711, Romania
tufis@racai.ro, elena@racai.ro

Abstract

The paper briefly describes the RoCo project and, in details, one of its first outcomes, the RoCo-News corpus. RoCo-News is a middle-sized journalistic corpus of Romanian, abundant in proper names, numerals and named entities. The initially raw text was previously segmented with MtSeg segmenter, then POS annotated with TNT tagger. RoCo-News was further lemmatized and validated. Because of limited human resources, time constraints and the dimension of the corpus, hand validation of each individual token was out of question. The validation stage required a coherent methodology for automatically identifying as many POS annotation and lemmatization errors as possible. The hand validation process was focused on these automatically spotted possible errors. This methodology relied on three main techniques for automatic detection of potential errors: 1. when lemmatizing the corpus, we extracted all the triples <word-form, POS tag, lemma> that were not found in the word-form lexicon; 2. we checked the correctness of POS annotation for closed class lexical categories, technique described by (Dickinson & Meurers, 2003); 3. we exploited the hypothesis (Tufiş, 1999) according to which an accurately tagged text, re-tagged with the language model learnt from it (biased evaluation) should have more than 98% tokens identically tagged.

1. Introduction

How important is to develop resources, and precisely corpora, is not a new issue for people involved in computational linguistic research. Being "a collection of pieces of language, selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language" (Sinclair, 1991), a corpus helps us to get a more comprehensive view of language in use. To Natural Language Processing researchers, corpora provide statistical evidence and realistic test beds.

RoCo-<domain> is a series of various registers corpora of Romanian that are developed within the Research Institute for Artificial Intelligence of the Romanian Academy and planned for public release to the research community. Currently, the automatically annotated corpora, not entirely validated, cover three registers: News, Literature and Legislation and consists of more than 35 millions of lexical tokens. All the corpora of the RoCo project are XML annotated and the minimal attributes for each lexical token are its morpho-lexical tag, and lemma. Additionally, the lexical tokens may be specified for hyphenation, may have a sense identifier and a chunk identifier (specifying which syntactic chunk the token belongs to).

In case of parallel corpora (which, currently, is the case for about 80% of the RoCo-<domain> corpora) sentence and word alignments are also included. A further extension will be the inclusion of dependency relations among the lexical tokens as well as anaphor-antecedent relations.

2. RoCo-News Corpus Description

A journalistic corpus is a good source of information about new words in a specific language vocabulary, about named entities, common abbreviations, and many other aspects of the functional style. In the following, we will describe RoCo-News corpus, representing, in its actual form, the result of a one-year validation and correction work of the automatic processing.

RoCo-News is a middle-sized journalistic corpus of Romanian. It contains about 7 million tokens, the number of distinct tokens being 231,626. The different articles in the corpus, initially available in various formats (doc, rtf and pdf) were converted into ASCII format with diacritical characters encoded as SGML entities. A preliminary text analysis revealed an abundance of proper names and number expressions. Specific to this kind of texts, the titles and authors names appear as distinct paragraphs and due to their partial grammatical structures may contain more tagging errors than the true paragraphs.

The raw text has been segmented with MtSeg segmenter developed in the context of the MULTEXT project (http://aune.lpl.univ-aix.fr/projects/Multext/), based on the segmentation resources developed in the MULTEXT-East project (http://nl.ijs.si/ME/), and POS tagged using the TnT tagger (Brants, 1998). The tagger has been trained on hand validated training corpus which includes the Orwell's "Ninety Eighty Four" novel (approx. 110,000 tokens), Plato's "Republic" (approx. 140,000 tokens) and several issues from some nation-wide newspapers (approx. 120,000 tokens). The tagset of the language model is derived from a large tagset, fully compatible with the MULTEXT-East morpho-lexical specifications (http://nl.ijs.si/ME/V3/msd/html/msd.html). The reduced tagset used throughout the RoCo-News corpus is the hidden tagset of the tiered tagging methodology (see (Tufiş, 1999) and (Tufiş & Dragomirescu, 2004) for further details). The reduced tagset contains 93 tags for words and 10 tags for punctuation.

The tagged corpus was further lemmatized. The lemmatization process is essentially a look-up procedure in a large word-form lexicon containing about 600,000 entries of the form: <word-form> <lemma> <tag>. The look-up procedure exploits the fact that knowing the tag of a word-form, its lemma is unique in the vast majority of cases (the few words - e.g. capete, capetele, copii, torturi, etc.- for which this is not the case, constitute an exception list, and their lemma is selected by statistical means). Since most of the words in a new text are likely to be present in the wide-coverage word-form lexicon, the lemmatization is highly accurate. In this step, the "unknown" marks assigned by the TnT tagger might be removed. The removal is conditioned by the correct tagging and its retrieval in the large word-form lexicon.

For the tokens not in this lexicon (and which are not tagged as proper names), the lemma is provided by our statistical lemmatizer. We use a set of rules (specific to each inflectional grammar category) automatically induced from the word-form lexicon that generate candidate lemmas for the unknown word and then Markov models (trained on lemmas from the lexicon) to rank the candidates. The one with the highest probability wins. This statistical lemmatization works very well, errors mostly happening when the unknown words belong to irregular inflection paradigms (Tufiş, 1989) or when their tagging was mistaken. Overall, considering the wide coverage of the word-form lexicon and the low error rate of the statistical lemmatizer, the probability of a lemmatization error is negligible.

3. Spotting possible errors in RoCo-News

All the tokens occurring in the RoCo-News unknown to the tagger (unseen during the training phase) were automatically marked by TnT with an asterisk.

We extracted from the RoCo-News corpus two files, one of them containing proper names, the other one listing <word-form lemma tag> triples that were not found in the word-form lexicon. They were used as key hints for discovering and (semi-automatically) solving a large number of tokenization and annotations errors (tags, lemmas or both). We also used the technique described in (Dickinson & Meurers, 2003) which spots possible errors by the analysis of the words belonging to close classes. Since errors correction is mostly manually done, the human factor might generate inconsistencies (not all instances of an error are corrected the same way, similar errors are dealt with differently or, simply, mistyping). As opposed to human annotation, the automatic tagging is much more consistent. Therefore we used the biased evaluation conjecture (Tufiş, 1999), which says that an accurately and consistently tagged text, re-tagged with the language model learnt from it (biased tagging) should reproduce almost identically (98%-99%) the original tagging. The differences are likely to spot most of the inconsistencies created by the human corrections as well as new errors unobserved before.

3.1. Lemmatization and re-tokenisation

Lemmatization is simpler and therefore more accurate than POS-tagging. In the vast majority of cases, the pair formed by a word occurrence and any of its legal tags would uniquely identify the lemma for that token if it is recorded into the word-form lexicon. However, the lemmatizer may produce wrong lemmas for unknown words, especially when the token is mistakenly tagged. The word-form lexicon is constantly updated as we find new entries in the texts we are working on. However, since this lexicon should be error-free, the new triples <word-form lemma tag> are subject to expert validation before being included into the word-form lexicon.

The lemmatization procedure is briefly described below:

a) If the current token is not marked by an asterisk and it was tagged by one of the following tags {AMPER, ASTER, COLON, DASH, DBLQ, DOLLAR, EXCL, EXCLHELIP, HELIP, LPAR, QUEST, QUOTE, PERIOD, RPAR, SCOLON, C, CR, I, M, R, Q, S, X} the occurrence form of the word is taken as its lemma. The rationale is that the token is either a punctuation token, or a word belonging to a non-inflectional grammar category.

b) If the current token is marked by the tagger as unknown, it is checked whether its POS annotation is NP, in which case the lemma is considered again being identical to the occurrence form of the token. The rationale is that in Romanian, proper names (foreign and male names) are rarely inflected. On the other hand, female Romanian names may be inflected, but the most frequent of them are already in the word-form lexicon.

The consecutive tokens tagged by the NP tag are concatenated and taken together as a single token the lemma of which is the concatenation of the respective lemmas. The left side in Table 1 exemplifies two sequences of proper nouns concatenated as shown in the right side of the table. Since the proper name Maria and its inflected forms are stored in the lexicon (it has no '*' mark-up) the lemma of the concatenated proper name Mariei_Ciupe is Maria_Ciupe.

…	…
soțul soț NSRY	soțul soț NSRY
Mariei Maria NP	Mariei_Ciupe Maria_Ciupe NP *
Ciupe Ciupe NP *	
este fi V3	este fi V3
şef şef NSRN	şef şef NSRN
la la S	la la S
Alpha Alpha NP *	Alpha_Bank Alpha_Bank NP*
Bank Bank NP *	
…	…

Table 1: Proper Names sequences and their concatenation (the asterisk marks the token unknown to the tagger

The unknown triples <word-form NP lemma> are saved in the *ProperNouns* file for later inspection and validation. The proper names in this file have been validated and the errors corrected. The corrections have been operated in the corpus as well. Some of the most typical errors were represented by tokens with all uppercase letters, or independent proper names being improperly concatenated.

The RoCo-News corpus contains many proper names that include middle name and/or first name abbreviation (e.g. Mircea M. Ionescu, M. D. Pavel) and in the majority of cases, their recognition as such was wrong (although the initials were always correctly tagged as abbreviations). The upper case initial(s) were automatically concatenated with the surrounding proper names and, after the hand validation, the correctly concatenated proper names (more than 99%) were added to the *ProperNouns* file (replacing whatever partial version of them).

The *ProperNouns* file, containing almost 30,000 distinct proper names, will be further extended with information concerning the type of the name entities (person, place, institution, etc.) the respective proper names denote.

c) If an unknown token is not tagged as NP, together with its tag, it is looked up in the word-form lexicon (which is much larger than the tagger's lexicon). If the pair <word-form tag> is found in the lexicon, its corresponding lemma is copied from the respective dictionary entry. Otherwise, the current token is processed by the probabilistic lemmatizer. In this case, the triple <word-form tag lemma> is saved in the file called *NotInTheLexicon* for subsequent inspection and validation.

The content of the *NotInTheLexicon* file was classified and analyzed in the decreasing order of the triples frequencies. It revealed more than 20,000 errors due to the wrong conversion of some diacritical characters into SGML entities or misspelling. This type of errors was systematic; once observed, it was simple to correct it.

Other major source of errors was the tokenization (misinterpretation of the period character, incomplete or incorrect specification of several frequent compounds, etc.). For instance, a fixed phrase (such as *de asemenea* "also", or *după-amiază* "afternoon"), which is specified into the tokenizer's resources will be concatenated. However, if in the training corpus the fixed phrase was not concatenated, it will be unknown to the tagger, and dealt with by the guesser.

A special case of unknown tokens is represented by the numbers. They are systematically tagged as numerals (M), but we noticed several cases where two or more consecutive numerals should have been taken together. For easier reading, it is customary to use a comma or a period among consecutive groups of three digits. However, we identified in our corpus several cases where the used separator was the blank and therefore the tokenizer considered the respective groups as distinct tokens. In this case, the distinct number tokens were concatenated similarly to proper names. Related to the number tokens, the analysis we conducted on the RoCo-News corpus outlined very regular patterns for associating more semantics to a number, useful for recognizing some named entity categories (amounts of money, dimensions, weights, telephone numbers, periods of time, ages, etc.). In Romanian it is rarely the case that a numeric NE is not accompanied by its specifier.

Among the unknown tokens, we found also web and e-mail addresses. These special tokens were systematically tagged as NN. We added to our tagset two new tags NNWEB and NNMAIL and all the occurrences of web and email addresses were retagged accordingly. The regular expressions describing such a token were included into the tokenizer.

3.2. Using closed class analysis for identifying errors

It is traditional in linguistics to divide lexical categories into two different types of classes: *closed classes* are those enumerable (e.g. classes like determiners, prepositions, modal verbs, or auxiliaries), whereas *open classes* are the large, productive categories such as verbs, nouns, adjectives. Dickinson and Meurers (2003) exploited the idea that, for detecting errors, one can make practical use of the closed-class concept. In the majority of the known tagsets, almost half of the tags correspond to closed-class categories of words. A closed-class category contains a reduced number of words, not very difficult to enumerate. Frequently, they can belong to different close-classes categories (e.g. in Romanian, one can find words that may be both prepositions and conjunctions or prepositions and auxiliaries, etc.). Depending on the tagset granularity, a closed-class category may accommodate several tags (e.g. various types of conjunctions, prepositions, pronouns, particles, etc.). Considering the fact that the closed-class words are very frequent, for a large corpus, one can safely assume that they occurred in most (if not all) of their possible contexts and thus it is reasonable to see all their different tags.

Dickinson and Meurers' idea was to search in a corpus for all occurrences of a closed-class tag and check whether each word is actually a member of his proper closed-class. We extracted from the word-form lexicon a list, L1, of closed-class tags, each of them indexing the set of words that could receive that tag. From this list, we computed another list, L2, containing words in L1 indexing two or more closed-class tags. Then, we extracted from the RoCo-News corpus all pairs <*word tag*> so that *tag* was a closed-class tag. If *word* was not in the set of L1 indexed by *tag* we checked the respective word occurrence in its context. In the majority of cases, we found a tagging error, but occasionally we also found errors in the word-form lexicon (a possible closed-class tag was not recorded for some words). Based on L2 we extracted all the words that were seen in the corpus only with a subset of their possible closed-class tagset. Some errors were again found in the word-form lexicon (words that were wrongly associated with some closed-class tags). We used a few regular expressions (defined in terms of surrounding tags for each target word) to extract sentences in which the not seen tags could have been licensed. Although this approach was not very precise (the most extracted sentences contained the correct tags), almost two hundred new errors were corrected.

Since there are still a few words in our word-form lexicon for which closed-class tags were not seen yet, we assume that the RoCo-news corpus may contain a (small) number of errors in the tagging of closed-class words.

3.3. Using biased evaluation for better error identification

The third technique used in cleaning up the RoCo-corpus was based on the biased evaluation conjecture (Tufiş, 1999) which says that an accurately and consistently tagged corpus, re-tagged with the language model learnt from it (biased evaluation), should have the vast majority of tokens identically tagged. The percentage of identical tags depends on the dimension of the corpus, but usually it is higher than 97.5%-98%.

After we made the corrections described in the previous sections, we took this version as the reference for the biased-evaluation procedure described in the following.

We trained the TnT tagger on the RoCo-News, building a new language model. We retagged RoCo-News with this new language model and compared the new tagging against the reference annotation. We found 96.8% identically tagged tokens and we extracted the differences.

Sorting the differences by their frequency, the first 100 difference types (accounting on average for 8-10,000 difference occurrences) were examined in context, one by one, and the validation expert decided which of the tags was correct (if any). Some of the differences were explained by inconsistent or partial corrections in the previous phases. Some other differences showed up because these corrections modified the contexts for the neighboring tokens and thus, according to the biased language model, many of them occurring in different contexts received different tags.

Correcting all the errors discovered in the analysis of the first 100 difference types ends the procedure. Given the dimension of the corpus, the procedure is very time consuming. We repeated this procedure several times with a continuous decrease of the number of differences. After months of analysis/correction cycles, the number of differences stabilized around 1.2% of the entire corpus.

At present, there are approximately 85,000 differences (22,500 distinct) between the reference RoCo corpus and its biased tagging version. The 1.2% differences are not evenly distributed. The analysis of the occurrences of the first 200 differences types, containing 168 different word-forms and accounting for about 18,929 differences, showed some surviving errors in the RoCo-News validated corpus, but also a reduced set of words which, although correct in the gold standard, were mistakenly tagged during the biased tagging.

The tag pairs appearing in the analyzed differences outlined 96 distinct confusion pairs. For each confusion pair we constructed the set of different words that were affected by the respective confusion. The more words affected by a confusion pair, the less worrying it is, because it might be ascribable to the inherent statistical noise. However, when a confusion pair is specific to a reduced number of words and if these words are frequent, it might be useful to have a closer look on the respective confusion pair. We made this investigation and discovered a few words that were responsible for significantly more differences than the rest.

Not surprising, the first four frequent words responsible for almost 3500 tag confusions are closed-class words. The weak forms of the personal pronouns (*le*, *ne, vă, îi*) show the highest error rate: out of 11,345 occurrences, 3,368 had wrong case label (accusative instead of dative and vice-versa). The correct case assignment for these pronouns is very hard when only distributional properties are taken into account (in general, the most frequent tag, i.e. the accusative one, will prevail). Deterministic treatment of the weak forms of pronouns would require information on the valency frame and (sometimes on) the sense of their main verbs. An alternative trivial solution to alleviate this problem is to drop the case distinction for the weak forms of the personal pronouns.

Another confusion, very difficult to avoid and relatively frequent in the RoCo-News corpus (210 occurrences), is *vor* tagged as main verb instead of auxiliary. This confusion is very unusual in Romanian because the main verbs and auxiliaries have different contexts. Moreover, no other word-form that could be either main or auxiliary verb (20) was affected by such confusion. The word *vor* (they want/they will) is quite frequent in the corpus (18,486 occurrences) and in the vast majority of cases (18,276) it is used as an auxiliary. The 210 wrong interpretations of it showed a regular pattern, (potentially much more productive): the word *vor* was followed by words that could be interpreted either as infinitives (as they should have been interpreted) or as nouns (as they were actually interpreted in the respective contexts). Since the only legal interpretation of *vor* when it is followed by the infinitive form of a verb is auxiliary (the construction is the future tense of the respective verb), the wrong tagging was the result of wrong tagging of the word following *vor*.

Some examples of words that generated the 210 wrong tagging of *vor* are the following: víza$_N$/vizá$_{Vinf}$, cífra$_N$/cifrá$_{Vinf}$, recólta$_N$/recoltá$_{Vinf}$, sărbătóri$_N$/sărbătorí$_{Vinf}$, táxa$_N$/taxá$_{Vinf}$, pláca$_N$/placá$_{Vinf}$, adrésa$_N$/adresá$_{Vinf}$, etc. For all occurrences of the homographic words that confused the tagging of the verb *vor,* the noun reading was much more frequent than the verb reading. This ambiguity does not exist in speech, because the homographs are not homophones being differentiated by the accent (shown in the examples by the accented vowel; for infinitives the accent is always on the final syllable).

4. Conclusion

We described a semi-automatic procedure by means of which we constructed a highly accurate annotated journalistic corpus for Romanian. Although it is language, tagger and tagset dependent, this approach is easy to apply for a different setting. The type of analysis we described gives strong indications about which words might be unreliably tagged. The procedure can be applied to other languages or other linguistic registers. In addition, it does not depend on the tagging method and does not require word-by-word inspection of the corpus. It does not ensure elimination of all existing errors, but the accuracy gain is substantial. We argued that repeating the biased tagging procedure and focusing only on the differences between the reference and the biased-tagged corpus, several types of errors can be removed and the difficulty to differentiate between morpho-lexical ambiguities can be easily spotted, allowing the corpus designer to make the appropriate decisions in order to solve them.

5. References

Brants, T. (1998). *TnT - A Statistical Part-of-Speech Tagger. Instalation and User Guide*, University of Saarland, Computational Linguistics, March 1998.

Brants, T. (2000). Inter-Annotated Agreement for a German Newspaper Corpus, In *Proceedings of LREC 2000.*

Dickinson, M., Meurers, W. D. (2003). Detecting Errors in Part of Speech Annotation. In *Proceedings of the 11th conference of the EACL-03*, Budapest, Hungary.

Sinclair, J. (1991). *Corpus, Concordance, Collocation,* Oxford University Press.

Tufiş, D. (1989). It Would Be Much Easier If WENT Were GOED. In Harry Somers, Mary McGee Wood (Eds.), *Proceedings of the 4th European Conference of the Association for Computational Linguistics,* Manchester.

Tufiş, D. (1999). Tiered Tagging and Combined Classifiers. In F. Jelinek, E. Nöth (Eds.) *Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence* 1692, Springer, 1999, pp. 28-33.

Tufiş, D., Dragomirescu, L. (2004). Tiered Tagging Revisited. In *Proceedings of the 4th LREC Conference.* Lisbon, Portugal, pp. 39-42.