The Lemmatisation Task at the EVALITA 2011 Evaluation Campaign

Fabio Tamburini

Dept. of Linguistics and Oriental Studies, University of Bologna, Italy
fabio.tamburini@unibo.it

Abstract. This paper reports on the EVALITA 2011 Lemmatisation task, an initiative for the evaluation of automatic lemmatisation tools specifically developed for the Italian language. Despite lemmatisation is often considered a subproduct of a PoS-tagging procedure that does not cause any particular problem, there are a lot of specific cases, certainly in Italian and in some other highly inflected languages, in which, given the same lexical class, we face a lemma ambiguity. A relevant number of scholars and teams participated experimenting their systems on the data provided by the task organisers. The results are very interesting and the overall performances of the participating systems were very high, exceeding, on interesting cases, 99% of lemmatisation accuracy.

Keywords: Evaluation, Lemmatisation, Italian.

1 Introduction

In the general linguistics view, lemmatisation is the process of grouping together the different inflected forms of a word so they can be analysed as a single item[1].

In computational linguistics, usually, two different methods are used to achieve this task: the first, called *stemming*, tries to reduce all the wordforms belonging to a specific paradigm to an invariant stem string, by removing all affixes, and does not, in general, produce a real surface string. The second, *lemmatisation*, identifies the process of transforming each wordform into its corresponding canonical base form, the lemma, where the canonical form is one particular wordform from an inflectional paradigm chosen by convention to represent the whole paradigm and, usually, corresponds to a headword found in a dictionary. In Italian, canonical base forms corresponds to verb inifinitives and the masculine singular form for nouns and adjectives (except for those cases that allow only the feminine gender).

Lemmatisation and stemming are normalisation techniques which proved to be very useful in a number of different NLP tasks, for information extraction and retrieval and to simplify corpus querying. The use of such normalisation methods helps automatic retrieval systems to remove wordform differences due to inflectional phenomena. They are both very relevant for highly inflected languages, for example romance, slavic and some northern european languages as well as a lot of other languages around the world, where the co-selection between bases and the different kind of affixes, both inflectional

[1] Collins English Dictionary, entry for ”lemmatise”.

B. Magnini et al. (Eds.): EVALITA 2012, LNCS 7689, pp. 230–238, 2013.

and derivational, can depend on a whole range of factors, from phonological to semantic (see [10] for a description of the different degree of inflection across languages).

In homograph handling we face essentially two types of ambiguities: *internal or grammatical ambiguities* when we encounter different wordforms belonging to the same lemma and consequently to the same part-of-speech (PoS) tag (e.g. *ami* as different forms of the verb *amare* - to love), and *external or lexical ambiguities* when considering wordforms belonging to different lemmas, but not necessarily to different PoS-tags (e.g. the verb form *perdono* in Table 1). Internal ambiguities do not matter for the lemmatisation task, because we should assign the same lemma, but for external ambiguities we face two very different cases: the first involves different PoS-tags and this is sufficient for choosing the correct lemma, but in the second case we can have two different lemmas presenting the same PoS-tag.

In the current literature, lemmatisation is often considered a subproduct of a PoS-tagging procedure that does not cause any particular problem. The common view is that no particular ambiguities have to be resolved once the correct PoS-tag has been assigned and a lot of the systems handling this task for different languages assume this view without indentifying and discussing the remaining potential external ambiguities [1,2,6,8,11,14], while some other scholars recognise the potential problem but ignore it [7].

Unfortunately there are a lot of specific cases, certainly in Italian and in some other highly inflected languages, in which, given the same lexical class, we face an external lemma ambiguity. The Table 1 shows some examples of such ambiguities for Italian. Homograph in verb forms belonging to different verbs or noun evaluative suffixation and plural forms are some phenomena that can create such kind of lemma ambiguities. A morphologically richer PoS-tagset could help alleviating the problem, at the price of a reduction in tagging accuracy, but in some cases the lemma ambiguity still persists.

Even the use of morphological analysers based on large lexica, which are undoubtedly very useful for the PoS-tagging procedures (see for example the results of the EVALITA2007 PoS-tagging task [12]), can create a lot of such ambiguities introducing more possibilities for creating homographs between different wordforms.

Certainly these phenomena are not pervasive and the total amount of such ambiguities is very limited, but we believe that it could be interesting to develop specific techniques to solve this generally underestimated problem.

2 Definition of the Task

The organisation provided two data sets: the first, referred to as Development Set (DS) contained a small set, composed of 17313 tokens, of data manually classified (see the following section for a detailed description) and were to be used to set up participants' systems; the second, referred to as Test Set (TS), contained the final test data for the evaluation and it was composed of 133756 tokens.

Lemmatisation is a complex process involving the entire lexicon. It is almost useless to provide a small set of training data for this task. No machine-learning algorithm would be able to acquire any useful information to successfully solve this task using only some hundred thousand annotated tokens. For these reasons, participants had to