EVALITA 2011: the Lemmatisation Task

Fabio Tamburini
DSLO - Università di Bologna
Email: fabio.tamburini@unibo.it

Outline

1. Motivation

- **Lemmatisation**: the process of transforming each wordform into its corresponding base form found in the dictionary (lemma).
- It is often considered a subproduct of a part-of-speech (PoS) 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.
- That is not always the case:

Wordform	PoS-tag	Possible Lemmas
cannone	NOUN	*cannone, canna*
piccione	NOUN	*piccione, piccia*
stazione	NOUN	*stazione, stazio*
morti	NOUN	*morto, morte*
aria	NOUN	*aria, ario*
macchina	NOUN	*macchina, macchia*
matematica	NOUN	*matematica, matematico*
osservatori	NOUN	*osservatore, osservatorio*
passano	VERB	*passare, passire*
danno	VERB	*dare, dannare*
perdono	VERB	*perdere, perdonare*

2. Data description

- **Document types**: journalistic and narrative genres, with small sections containing academic and legal/administrative prose.
- **Data sets**: Development Set (DS) = 17,313 tokens, Test Set (TS) = 133,756 tokens, Ratio between DS and TS is 1/8. Both sets were manually annotated (PoS-tags and lemmas).
- **Additional resources**: Lemmatisation is a complex process involving the entire lexicon. It is almost useless to provide a small set of training data for this task. For these reasons, participants were allowed to use other resources in their systems, both for develop and to enhance the final performances.
- **PoS-Tagset**: we used a "traditional" tagset (EAGLES-like), the same used in the EVALITA 2007 PoS-tagging task.
- **Tokenisation**: All the development and test data were provided in tokenised format, one token per line followed by its tag.

3. Evaluation metric

- The evaluation was performed in a "black-box" approach: only the systems' outputs were evaluated.
- The evaluation metric was based on a token-by-token comparison and only one lemma was allowed for each token.
- The evaluation was only referred to open class words and not to functional words: only the tokens having a PoS-tag comprised in the set {ADJ_*, ADV, NN, V_*} had to be lemmatised.
- The considered metric was: **Lemmatisation Accuracy**, defined as the number of correct lemma assignments divided by the total number of tokens in the TS belonging to the lexical classes considered for the evaluation (65210 tokens).

4. Participating systems

Four systems completed all the steps in the evaluation procedure and their outputs were officially submitted for this task by their developers.

Research Team	Affiliations	System Label
Rodolfo Delmonte	University of Venice, Italy	Delmonte_UniVE
Djamé Seddah	Alpage (Inria)/Univ. Paris Sorbonne, France	Seddah_Inria-UniSorbonne
Maria Simi	University of Pisa, Italy	Simi_UniPI
Fabio Tamburini	University of Bologna, Italy	Tamburini_UniBO

Universities
Research Institutions
Companies

Systems Descriptions

- **Delmonte_UniVE** - a rule based lemmatiser based on a lexicon composed of about 80.000 lemmas and additional modules for managing ambiguities based on frequency information extracted from various sources.
- **Seddah_Inria-UniSorbonne** - a tool for supervised learning of inflectional morphology as a base for building a PoS-tagger and a lemmatiser and a lexicon extracted from Morph-It and the Turin University Treebank.
- **Simi_UniPI** - a basic lemmatiser based on about 1.3 millions of wordforms followed by a cascade of filters (affix specific management, search in Wikipedia or directly on Google for similar contexts, ...).
- **Tamburini_UniBO** - a morphological analyser based on Finite State Automata equipped with a large lexicon of 110.000 lemmas and a simple algorithm that relies on the lemma frequency classification proposed in the De Mauro/Paravia dictionary.

5. Results

Four, very simple and naïve, baseline systems were introduced by the organisers:

- **Baseline_1**, simple copy,
- **Baseline_2**, V_ESSERE or V_AVERE and V_MOD corrections,
- **Baseline_3** De Mauro/Paravia online dictionary + Levenshtein distance,
- **Baseline_4** searches into the DS lexicon for a reference lemma.

SYSTEM	LEMMATISATION ACCURACY
Simi_UniPI	**99.06**%
Tamburini_UniBo	98.74%
Delmonte_UniVE	98.42%
Seddah_Inria-UniSorbonne	94.76%
Baseline_4	83.42%
Baseline_3	66.20%
Baseline_2	59.46%
Baseline_1	50.27%

6. Discussion

Examining the systems' performances with respect to their structural features, we can make some tentative observations:

- The results were quite hight, mostly of them above 98%. Considering that only half of the total number of tokens in the TS have been evaluated, these results depict a good global picture for this evaluation task.
- The neat separation between the baselines performances and the real systems can suggest that this task cannot be solved by using simple heuristics, but the disambiguation process has to be based on various sources of information: large lexica, frequency lists, powerful lemmatiser morphology-aware and so on.
- Only the best performing system, in our knowledge, use the sentence context to choose among the different lemmas connected to an ambiguous wordform. Maybe this could be the most promising direction for increasing the automatic system performances for the lemmatisation task.

THANK YOU!

and

THANKS TO ALL TASK PARTICIPANTS AND EVALITA ORGANISERS!