EVALITA 2011

The Tanl Lemmatizer Enriched with a Sequence of Cascading Filters

G. Attardi, S. Dei Rossi, M. Simi
Dipartimento di Informatica – Università di Pisa

IN SUPREMÆ DIGNITATIS
1343

Approach

- The base Tanl PoS tagger and Lemmatizer
  - Rich tagset with morpho-syntactic information
  - A large Italian lexicon (1.2 million forms)
- Using fine grained PoS reduces ambiguities
  - Less than 2400 ambiguities in full-form lexicon

danno	VERB	Vip1s	dannare	Vip3p	dare
cannone	NOUN	Sms	cannone	Sfp	canna

- Tanl lemmatizer enriched with cascading filters
  - Forms in lexicon
  - Unknown words: morphological alterations, prefixes …
  - Semantic disambiguation based on *deep* Wikipedia index and Google

The Tanl PoS tagger

- The Tanl PoS tagger
  - Tanl tagset (336 *morphed* tags)
  - Tanl tagger (derived from Tree Tagger: added memory mapping and UTF-8 support)
  - Basic lemmatizer (no strategy for unknown words; first lemma)
- Italian lexicon
  - Base lexicon (65,650 forms), full-form lexicon (1,273,200 forms); inflection rules supplied by Achim Stein and extended
  - Aligned to the Tanl tagset

The Filters Architecture

- **1st Filter: Word in Lexicon**

```
If morphed pos compatible with the gold pos
  return the lemma (or lemmas) associated to the morphed pos
else return the lemma (or lemmas) compatible with the gold coarse pos
```

- **2nd Filter: Morphological Alterations**
  - Using a suffix list, try to rewrite morphological alteration of words such as augmentative, diminutive, depreciative, terms of endearment …
- **3rd Filter: Check for the Existence of Common Prefixes**
  - *anti*, *pre*, *ri, auto* … , we try to lemmatize the form obtained by trimming the prefix;
- **4th Filter: Guess Lemma**
  - Try to generate the lemma by using a list of common suffixes, if unable use the form as lemma
- **5th – 6th Filters: resolving lemmas ambiguities**

Deep search on Wikipedia

- Search engine built on Wikipedia, which exploits syntactic and semantic annotations added to the Italian Wikipedia texts by the Tanl linguistic pipeline [SemaWiki project]
  - word form, PoS tag, lemma
  - NE category, super sense
  - dependency information (result of the DeSR dependency parser)
- Possible queries

– *Chi è Cleopatra?*	DEP/subj:Cleopatra MORPH/essere:*
– *Chi ha ucciso Cesare?*	deprel [DEP/obj:Cesare MORPH/uccidere:*

Semantic disambiguation

- AskWiki
  - “Deep Search” on the Italian Wikipedia
  - Given noun “pupille”, lemma is “pupilla” or “pupillo”?
    - MORPH/iride:* pupilla: 27 hits
    - MORPH/iride:* pupillo: 0 hits
- AskGoogle (if still unresolved)
  - Given noun “conti”, lemma is “conto” or “conte”?
    - "accreditamento * conto" : 51600 hits
    - "accreditamento * conte" : 2 hits
  - Limit: 100 queries per day

Breakdown of results

Stages	Accuracy	Improvement
Task baseline (version 4)	83.42%	
Our baseline	96.65%	30.45 %
1st – Word in Lexicon	98.48%	1.83 %
2nd – Morphological Alterations	98.60%	0.12 %
3rd – Common Prefixes	98.61%	0.01 %
4th – Guess Lemma	98.98%	0.37 %
5th – askWiki	99.05%	0.07 %
6th – askGoogle	**99.06%**	0.01 %

Error analysis

~500 errors on the test set

Error type	**Percentage**
Errors in guessing nouns and adjectives	33.9 %
Errors in dealing with alterations	24.8 %
Errors in guessing verbs	10.2 %
Errors in resolving ambiguities	*9.3 %*
Errors in dealing with truncated words	8.5 %
Errors in dealing with clitics	4.9 %
Errors in the gold test	3.9 %
Lexicon differences w.r.t. task conventions	1.8 %
Foreign words	1.6 %
Failures in dealing with prefixes	1.0 %

Conclusions

- Task was useful in
  - Improving the lexicon (after task we achieved 99.53% accuracy)
  - Highlighting that simple strategies for unknown words may be effective
- Using finer PoS tags can greatly reduce alternative lemmas
  - Genuine semantic ambiguities account for less than 10% of errors
  - Resorting to external resource is costly and may not be worthwhile
- Future work: give priorities to alternatives