Morphological Analysis Without Expert Annotation

Garrett Nicolai and Greg Kondrak

UNIVERSITY OF ALBERTA

EDMONTON·ALBERTA·CANADA

1. Introduction and Motivation

- The task of morphological analysis is to annotate a given word-form with its lemma and morphological tag.
- A single word-form may have several correct analyses (possible inflections for German "lüfte" are on the right).
- Lexicons and Finite-State Analyzers are expensive to create.
- Unlike Morphological tagging, Morphological analysis is context-free.
- We propose a method that trains on inflection tables, rather than morphologically annotated corpora.
- We are more accurate than the analysis module of a morphological tagger, with considerably higher coverage than an FST analyzer.

Lemma	Inflection	Tag
Luft	Nom Pl (Noun)	NP
Luft	Acc Pl (Noun)	AP
Luft	Gen. Pl (Noun)	GP
lüften	1st Sg Ind Pres (Verb)	1SIE
lüften	1st Sg Sub Pres (Verb)	1SKE
lüften	3rd Sg Sub Pres (Verb)	3SKE
lüften	Imperative Sg (Verb)	RS

2. Methods

2.1 Alignment

- Words are aligned to lemma+tag representations using m2m aligner.
- Tags are indivisible, and align to a single affix.

l	L
ü	u
f	f
t	t
e	+
	NP

2.2 Transduction

- After alignment, extract and weight transduction rules using DirecTL+:
  - Context: ü → u / ^l_fte$
  - Markov: e → +NP / uft_$
  - LC: e→+NP / üft_$ && uft_$
  - Joint: e →+NP / ü:u f:f t:t _ $
  - Copy: f → f
- Produce *n*-best list of predictions.
- Re-rank and threshold.

2.3 Re-ranking

Features

- Does lemma occur in raw corpus?
- Normalized character LM score.
- Normalized transduction score.
- Affix-match: was aligned affix / tag pair seen in training?
- Mirror constraint: does putting the proposed analysis through a generator produce the word? (ie, g(f(w)) = w)

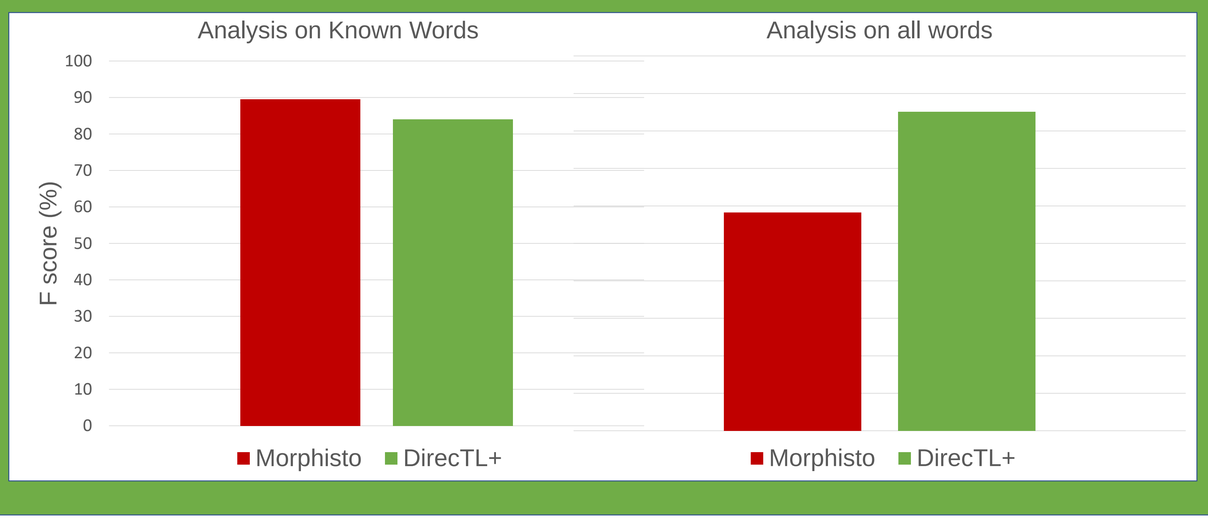
Mirror Constraint

Word	Analysis	Mirror
lüfte	luften+1SIE	lufte
lüfte	loften+3SKE	löfte
lüfte	lüften+3SKE	lüfte
lüfte	lüften+3SIA	lief
lüfte	lüften+3SIE	lüftet

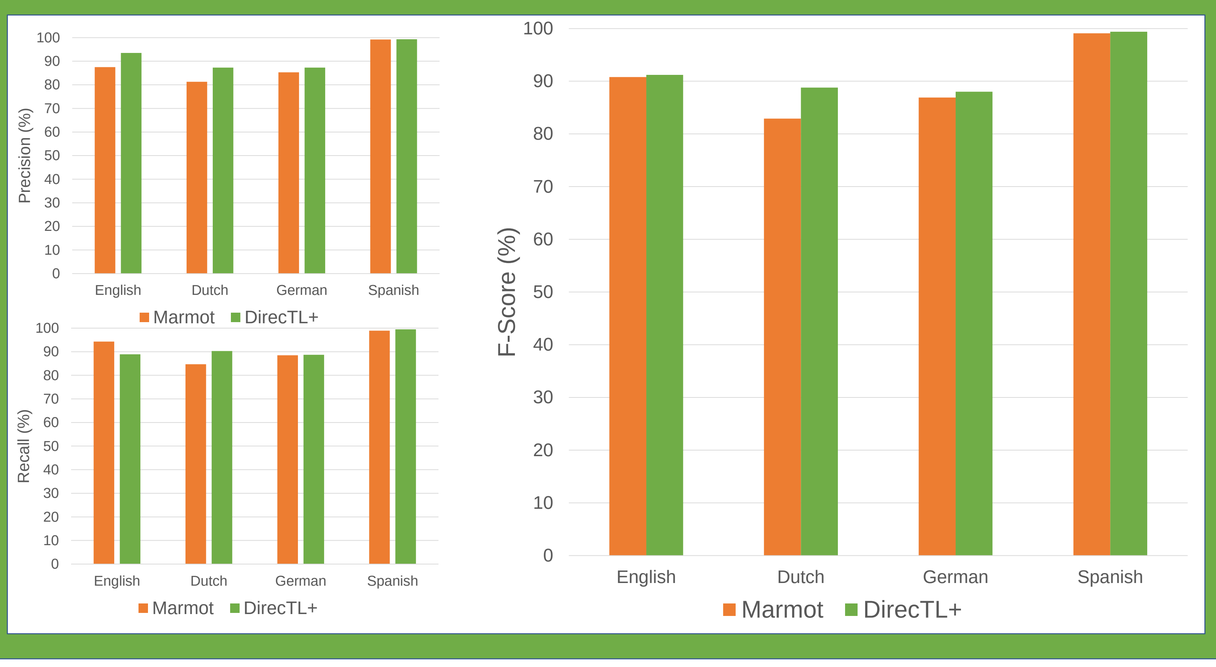
3. Experiments

- Data is from CELEX (English, Dutch, German), and Wiktionary (Spanish):
  - English, Dutch, German: verbs, nouns, adjectives
  - Spanish: verbs
- We evaluate micro-averaged German F-score against Morphisto, an FST analyzer.
- We evaluate macro-averaged F-score against Marmot.
  - Both systems make use of unannotated Wikipedia corpora.

3.2 Comparison against FST (German)



3.3 Comparison against Marmot



4. Conclusion

- Our method approaches the accuracy of a hand-crafted morphological analyzer, but has much higher coverage.
- Our method is also more accurate than the analysis module of a state-of-the-art morphological tagger.
- Access to inflection tables reduces the need for expert-crafted morphological lexicons.

Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council of Canada, and Alberta Innovates Technology Futures.