DerivBase.hr

LREC 2014 @ Reykjavik

A High-Coverage Derivational Morphology Resource for Croatian

Jan Šnajder, TakeLab, FER, University of Zagreb

http://takelab.fer.hr/derivbasehr

about

What?

A resource that groups **100K Croatian lemmas** into **~50K clusters** of **derivationaly related lemmas** (deriv. families)

Why?

Knowledge about derivational morphology is useful for various NLP tasks: semantic similarity, textual entailment, SRL

How?

Following the approach of **German DErivBase** [1], we induce the derivational families from corpus using a **rule-based framework for modeling deriv. morphology**.
Focus is on **suffixal derivation** of nouns, adjectives, and verbs

Challenges

(1) Obtaining a clean and comprehensive list of lemmas
(2) Comprehensive modeling of derivation
(3) Resource evaluation

model

Uses Higher-Order Functional Morphology (HOFM) framework [2]

(1) Inflectional component

93 manually defined inflectional paradigms for nouns, verbs, and adjectives. Succinct representation of complex morphological transformations (stem changes, optionality, etc.)

(2) Derivational component

244 manually defined suffixal derivational patterns between pairs of inflectional patterns. E.g.:

$d = (t, I_1, I_2) = (sfx(\text{“ica”}) \circ try(jat) \circ try(plt), Nf, Nf)$
$svijeća_{N01}$ *(candle)* → $svjećica_{N02}$ *(small candle)*
$kuća_{N01}$ *(house)* → $kućica_{N02}$ *(small house)*

Patterns define the admissible derivations and thus routinely **overgenerate**

$L_d(l, p) = \{(l_1, p_1), \ldots, (l_n, p_n)\}$
$L_d(kuća_{N01}) = \{kućica_{N02}, *kućica_{N03}\}$
$L_d(bol_{N03}) = \{*bolica_{N02}\}$

To remove spurious derivations, we **filter** against an inflectional lexicon acquired from corpus

construction

Step 1: Corpus preprocessing

Corpus is a 12BW **hrWaC** [3], POS tagged and lemmatized
We extract **1.2M lemma-POS pairs** (POS={N,A,V})
Insufficient quality: only 16% of lemma-POS pairs are correct

Step 2: Inflectional lexicon acquisition

(1) Choose the most plausible paradigm for each lemma-POS
(most plausible = produces most corpus-attested wordforms)
(2) Remove overlapping lemma-paradigms (false homographs)
Results in **100K lemma-paradigm pairs** (42.3% F1-score)

Step 3: Cluster induction

(U) Unsupervised method:
Hierarchical agglomerative clustering based on a suffix-sensitive string-distance measure
Results in **38K clusters**
(K) Knowledge-based method:
Equivalence classes of the derivational relation induced by the derivational patterns D:
$(l_1, p_1) \rightarrow_D (l_2, p_2)$ *iff* $\exists d \in D.\ (l_2, p_2) \in L_d(l_1, p_1)$
Results in **56K clusters**

evaluation

Gold standard

Step 1: Acquired **sample of 50 complete deriv. families**
Step 2: Sampled **2000 pos and 2000 neg lemma pairs**
(pos = in the same DF, neg = string-similar but not in the same DF)
Step 3: Manual annotation of lemma pairs
pos: **R** (deriv. + sem. related), **M** (deriv. related)
neg: **N** (no relation), **L** (lemmatisation error), **C** (composition)

Results

	#clusters	P	R	F_1
DerivBase.hr (U)	37,999	76.0	75.4	75.7
DerivBase.hr (K)	55,551	**81.2**	**76.5**	**78.8**
Prefix stemmer	62,228	49.2	42.1	45.4
Rule-based stemmer	93,098	25.0	0.4	0.9

Future improvements

Recall: (1) new patterns, (2) pattern compositions
Precision: (1) predict derivational relation, (2) clustering
Both: improve inflectional lexicon acquisition

refs

[1] B. Zeller, J. Šnajder, S. Padó (2013). DErivBase: Inducing and evaluating a derivational morphology resource for German. ACL 2013, pp. 1201–1211.
[2] J. Šnajder and B. Dalbelo Bašić (2010). A computational model of Croatian derivational morphology. FASSBL 2010, pp. 109–118.
[3] N. Ljubešić and T. Erjavec (2011). hrWaC and slWac: Compiling web corpora for Croatian and Slovene. TSD 2011, pp. 395–402.

Text Analysis and Knowledge Engineering Lab
Faculty of Electrical Engineering and Computing
University of Zagreb, Croatia

TakeLab FER

Work supported by the **Croatian Science Foundation**
Grant 02.03/162: *Derivational Semantic Models for IR*