DERIVBASE.HR: A High-Coverage Derivational Morphology Resource for Croatian

Jan Šnajder
University of Zagreb, Faculty of Electrical Engineering and Computing
Text Analysis and Knowledge Engineering Lab
Unska 3, 10000 Zagreb, Croatia
jan.snajder@fer.hr

Abstract

Knowledge about derivational morphology has been proven useful for a number of natural language processing (NLP) tasks. We describe the construction and evaluation of DERIVBASE.HR, a large-coverage morphological resource for Croatian. DERIVBASE.HR groups 100k lemmas from web corpus hrWaC into 56k clusters of derivationally related lemmas, so-called *derivational families*. We focus on suffixal derivation between and within nouns, verbs, and adjectives. We propose two approaches: an unsupervised approach and a knowledge-based approach based on a hand-crafted morphology model but without using any additional lexico-semantic resources. The resource acquisition procedure consists of three steps: corpus preprocessing, acquisition of an inflectional lexicon, and the induction of derivational families. We describe an evaluation methodology based on manually constructed derivational families from which we sample and annotate pairs of lemmas. We evaluate DERIVBASE.HR on the so-obtained sample, and show that the knowledge-based version attains good clustering quality of 81.2% precision, 76.5% recall, and 78.8% F_1-score. As with similar resources for other languages, we expect DERIVBASE.HR to be useful for a number of NLP tasks.

Keywords: derivational morphology, lexical resource, Croatian language

1. Introduction

Morphological processing is a prerequisite for many natural language processing (NLP) applications. Most work focuses on inflectional morphology and tasks such as lemmatization, part-of-speech (POS) tagging, and paradigm induction. In contrast, morphological derivation, which describes the creation of new words from the existing ones, has received far less attention. Derivation is especially interesting for morphologically complex languages, such as the Slavic languages, which have a very productive derivational morphology. Knowledge about derivational morphology has been proven useful for a number of NLP tasks, ranging from semantic similarity (Luong et al., 2013; Padó et al., 2013) to textual entailment (Shnarch et al., 2011) and semantic role labeling (Green et al., 2004).

In this paper we describe the induction and evaluation of DERIVBASE.HR, a large-coverage morphological resource for Croatian, which groups lemmas from corpus into clusters of derivationally related lemmas, so-called *derivational families*. We follow a procedure similar to the one employed for German by Zeller et al. (2013). However, as Croatian is more complex with regard to both inflection and derivation, the task is more challenging.

We focus on the very productive suffixal derivation between and within nouns, verbs, and adjectives. We propose two approaches: an unsupervised approach and a knowledge-based approach based on derivational patterns but without using any additional lexico-semantic resources. We perform an evaluation on a manually annotated sample. The resulting resource has a high-coverage (98K lemmas) and good quality (80.8% precision and 76.1% recall). We make the resource freely available.[1]

[1] http://takelab.fer.hr/derivbasehr

2. Related Work

There are a number of resources targeting morphological derivation. WordNet has been extended to include *morphosemantic* relations for English (Fellbaum et al., 2009) and a number of other languages (Bilgin et al., 2004; Pala and Hlaváčková, 2007; Koeva et al., 2008; Šojat and Srebačić, 2014). CELEX database (Baayen et al., 1996) provides derivational knowledge for English, German, and Dutch. CatVar (Habash and Dorr, 2003) is a resource aimed specifically at derivation, which groups English nouns, verbs, and adjectives into derivational families. CatVar has been used in many applications, including paraphrase detection and semantic role labeling.

More recently, Zeller et al. (2013) constructed DERIVBASE, a large-coverage derivational resource for German covering 280k lemmas from a web corpus. Their approach relies on clustering based on hand-crafted derivational rules, and achieves 93% precision and 71% recall.

In this work, we follow the approach of Zeller et. al, but also address a number of issues arising from the morphological complexity of Croatian language. First, because Croatian language is more complex than German regarding inflection, we have to rely on a much more elaborated model of inflection. Thus, unlike Zeller et. al, who use a set of lemmas and their parts of speech as input to the clustering method, we have to obtain a set of lemma-paradigm pairs, which is arguably a more difficult task. Secondly, even when restricted to suffixal derivation, Croatian derivational morphology is still more complex than German: Zeller et al. report implementing 158 derivational patterns, while we use a model that consists of 244 patterns. This makes it harder to attain clusters with high recall.

A related area of research is that of unsupervised morphology (Hammarström and Borin, 2011). Gaussier (1999) builds French derivational families using an unsupervised

method. Piasecki et al. (2012) bootstrap derivation rules starting from seed examples. In most cases, however, unsupervised approaches are not able to differentiate between inflectional and derivational morphology. In this work we experiment with an unsupervised method based on string distance-based clustering, but demonstrate that it performs worse than a knowledge-based approach.

There exist a number of computational morphology resources for the Croatian language, including inflectional morphological lexica (Tadić and Fulgosi, 2003; Šnajder et al., 2008) as well as morphological analyzers and lemmatizers (Ćavar et al., 2009; Agić et al., 2013). There has also been work on computational models of derivational morphology: Šnajder and Dalbelo Bašić (2010) present a computational model of Croatian suffixal derivation, while Šojat and Srebačić (2014) analyze the morphosemantic relations between Croatian verbs and discuss their inclusion in Croatian WordNet. However, there appears to be no prior work on inducing a large-coverage derivational morphology resource such as DERIVBASE.HR.

3. Morphology Model

Our knowledge-based approach to DERIVBASE.HR induction relies on a generative model of morphology. Similarly to Zeller et al. (2013), we use the HOFM modeling framework proposed by Šnajder and Dalbelo Bašić (2008) and Šnajder and Bašić (2010). The model of Croatian morphology consists of an inflectional and a derivational component; we use the former to obtain the lemmas and the latter for knowledge-based induction of clusters. Both components are freely available.[2] For details, the reader is referred to (Šnajder and Dalbelo Bašić, 2008; Šnajder and Bašić, 2010; Šnajder, 2010).

3.1. Inflectional Component

The inflectional component defines the inflectional paradigms for nouns, verbs, and adjectives. The current version of the model uses 93 paradigms. The HOFM formalism uses a succinct representation of string-based transformations, allowing for compact representation of more complex inflectional paradigms, and phenomena such as stem changes and optionality. Table 1 lists some example transformations.

3.2. Derivational Component

The basic building blocks of the derivational component are the *transformation functions* and *derivational patterns*. A derivational pattern describes the derivation of a derived word from a basis word. A derivational pattern is a triple

$$d = (t, \mathcal{I}_1, \mathcal{I}_2)$$

where t is the transformation function that maps the word's stem (or lemma) into the derived word's stem (or lemma), while $\mathcal{I}_1$ and $\mathcal{I}_2$ are the sets of inflectional paradigms of the basis word and the derived word, respectively. The transformation of the lemma into the stem and vice versa is handled by the underlying inflectional component, depending on the inflectional paradigm associated with the lemma.

[2] http://takelab.fer.hr/hofm

Function	Description
$sfx(s)$	concatenate the suffix s
$dsfx(s)$	delete the suffix s
$aifx(s1, s2)$	alternate the infix $s1$ to $s2$
$try(t)$	perform transformation t, if possible
$opt(t)$	optionally perform transformation t
plt, jot , jat	alternate infixes for palatalization/jotation/jat
Examples	
$\big(sfx(\text{“}en\text{”}) \circ opt(jot), \mathcal{N}_{mf}, \mathcal{A}_q\big)$ “Derive qualificative adjective from a noun by suffixation of *-en* and optional jotation	
$\big(sfx(\text{“}ica\text{”}) \circ try(jat) \circ try(plt), \mathcal{N}_f, \mathcal{N}_f\big)$ “Derive diminutive feminine *-ica* nouns from feminine nouns, if possible with jat reflex alternation and palatalization”	

Table 1: Transformation functions and exemplary derivational patterns in the framework by Šnajder (2010)

Note that each derivation pattern defines the admissible derivations, some of which may not exists in the language. Also, patterns say nothing about the semantic relation between the basis and the derived word. Otherwise said, the derivational patterns tend to overgenerate. Additional filtering based on corpus-attested lemmas ensures that overgeneration is kept to a minimum.

The model of Croatian derivational morphology currently consists of 244 suffixal derivational patterns, which describe the derivation of nouns, verbs, and adjectives. The model covers optional transformations and phonologically conditioned alternations of the stem and suffix. Table 1 shows examples of derivational patterns.

4. Inducing the Resource

The induction of DERIVBASE.HR from corpus consists of three steps: (1) corpus preprocessing, (2) acquisition of an inflectional lexicon, and (3) the induction of derivational clusters.

4.1. Corpus Preprocessing

As a stating point for DERIVBASE.HR induction, we use hrWaC, the 12B-token Croatian web corpus compiled by Ljubešić and Erjavec (2011). For POS-tagging and lemmatization, we use the tools developed by Agić et al. (2013), based on the HunPos tagger (Halácsy et al., 2007) and the CST lemmatizer (Ingason et al., 2008). The accuracy of the tagger and lemmatizer on newspaper corpora is 97% and 98%, respectively. As reported by Šnajder et al. (2013), tagging and lemmatization accuracy on a mixed-domain corpus (Wikipedia) drops to about 94% and 96%, and we can expect further decreases on hrWaC due to lower linguistic quality. Thus, to remove the invalid lemmas, we first filter out all lemmas not tagged as either a noun, adjective, or verb, as well as lemma-POS pairs that occur less than tree times in the corpus. After filtering, we end up with 1.2M lemma-POS pairs, the majority of which still are invalid. We sampled 200 lemma-POS pairs and manually annotated their validity: only 16% lemma-POS pairs were valid (we consider a lemma-POS to be valid if both the lemma and the POS are correct, the lemma is not a proper

name, nor is it derived from a proper name). Obviously, the list of lemmas must be further filtered out to obtain a good quality resource.

4.2. Acquiring the Inflectional Lexicon

As outlined in Section 3.2., the derivational component requires each lemma to be associated with its inflectional paradigm. The set of lemmas associated with their paradigms constitutes an *inflectional lexicon*. Thus, the next step in building of DERIVBASE.HR is the acquisition of an inflectional lexicon from hrWaC. This serves two purposes: filtering of invalid lemmas and setting up the ground for knowledge-based induction of derivational clusters.

The task of lexicon acquisition has been extensively studied in the literature, e.g. (Oliver, 2003; Sagot, 2005; Hana, 2008; Šnajder et al., 2008). In our case, the task is simpler because we start out with a list of lemma-POS pairs, although an imperfect one.

We employ a two-step procedure: for each lemma-POS pair, we consider all applicable inflectional patterns and choose the most plausible one. The most plausible inflectional pattern is the one that produces the most corpus-attested wordforms above a specified threshold (we set this threshold to 4 wordform types and 10 wordform tokens). In the second step, we remove heuristically the lemma-paradigm pairs for which the overlap of corpus-attested wordforms exceeds a specified threshold (we set this threshold to 3). The acquired inflectional lexicon $\mathcal{L}$ contains 100k lemmas (58.5k nouns, 29.5k adjectives, and 12k verbs). When evaluated on the sample of 200 lemma-POS pairs, the lexicon achieves 55.0% precision, 34.4% recall, and 42.3% F_1-score. This performance could be further improved by optimizing the acquisition parameters, but we leave this for future work.

4.3. Unsupervised Induction

For unsupervised induction of DERIVBASE.HR, we resort to clustering of lemmas based on a string-distance measure. We use the measure proposed by Majumder et al. (2007), which in several studies (Šnajder and Dalbelo Bašić, 2009; Zeller et al., 2013) has show to be effective in capturing suffixal variation. For words X and Y, it is defined as

$$D_4(X,Y) = \frac{n-m+1}{n+1} \sum_{i=m}^{n} \frac{1}{2^{i-m}} \qquad (1)$$

where m is the position of left-most character mismatch, and $n+1$ is the length of the longer of the two strings. Notice that this is a suffix-oriented measure; for measuring prefix variation, one can apply the measure on the reversed string.

For the actual clustering we use the hierarchical agglomerative algorithm with average linkage. The space complexity of this algorithm is quadratic and prohibits clustering of all lemmas at once. Instead, we precluster by recursively partitioning the set of lemmas sharing the same prefix into partitions of 1000 lemmas, which we then cluster separately. The number of clusters is optimized for F1-score on the development set (cf. Section 5.2.). The resulting resource contains 38k derivational clusters (avg. of 2.6 lemmas per cluster), of which 22k (58%) are non-singleton clusters.

(1)	*razljutiti*$_V$ *razljućen*$_A$
(2)	*razložen*$_A$ *razložnost*$_N$ *razlog*$_N$ *razložan*$_A$ *razložiti*$_V$
(3)	*razlomiti*$_V$ *razlomljen*$_A$ *razlomak*$_N$
(4)	*razlupan*$_A$
(5)	*razlučivost*$_N$ *razlučan*$_A$ *razlučiti*$_V$ *razlučiv*$_A$ *razlučivati*$_V$ *razlučivanje*$_N$ *razlučen*$_A$
(6)	*razmaknica*$_N$ *razm*$_N$ *razmak*$_N$ *razmiti*$_V$
(7)	*razmahivati*$_V$ *razmahan*$_A$ *razmah*$_N$ *razmahati*$_V$
(8)	*razmaknuti*$_V$ *razmaknut*$_A$
(9)	*razmatati*$_V$ *razmatanje*$_N$
(10)	*razmatrati*$_V$ *razmatranje*$_N$ *razmatran*$_A$ *razmazivati*$_V$ *razmaziti*$_V$ *razmazan*$_A$ *razmažen*$_A$ *razmazivanje*$_N$ *razmaz*$_N$ *razmazati*$_V$
(11)	*razmekšati*$_V$
(12)	*razmetljivac*$_N$ *razmetati*$_V$ *razmetljiv*$_A$ *razmetanje*$_N$ *razmetan*$_A$

Table 2: An excerpt of 12 clusters from DERIVBASE.HR

4.4. Knowledge-Based Induction

For knowledge-based induction, we use the inflectional lexicon $\mathcal{L}$ and a set derivational patterns to induce the derivational families. Given a lemma-paradigm pair (l,p) as input, a single derivational pattern $d = (t, \mathcal{I}_1, \mathcal{I}_2)$ generates a set of possible derivations $L_d(l,p) = \{(l_1,p_1),\ldots,(l_n,p_n)\}$, where $p_i \in \mathcal{I}_1$ and $p_i \in \mathcal{I}_2$. Given a set of derivational patterns $\mathcal{D}$, we define a binary derivation relation $\rightarrow_{\mathcal{D}}$ between two lemma-paradigm pairs that holds if the second pair can be derived from the first one as:

$$(l_1,p_1) \rightarrow_{\mathcal{D}} (l_2,p_2) \quad \text{iff} \quad \exists d \in \mathcal{D}.\, (l_2,p_2) \in L_d(l_1,p_1)$$

We now define derivational families as the equivalence classes of the transitive, symmetric, and reflexive closure of $\rightarrow_{\mathcal{D}}$ over $\mathcal{L}$. Note that transitivity is achieved only over the entries in $\mathcal{L}$, so low coverage of $\mathcal{L}$ will result in fragmented derivational families. Conversely, because derivational patterns overgenerate, $\mathcal{L}$ should not contain many invalid lemmas.

After computing the equivalence classes, we evaluated the resource on the development set (cf. Section 5.2.). To improve the recall, we implemented additional 30 derivational patterns. The resulting resource contains 56k derivational clusters (avg. of 1.8 lemmas per cluster), of which 15k (27%) are non-singleton clusters. Table 2 shows an excerpt from the resource.

5. Evaluation Methodology

The induction of derivational families is essentially a clustering task and could be evaluated as such. While a number of clustering evaluation metrics have been proposed (see (Amigó et al., 2009) for an overview), there is no consensus on the best approach. Moreover, due to semantic drifts involved with derivation, building a representative gold standard for cluster-based evaluation of derivational families is in itself not a straightforward task. Thus, instead of performing a cluster-based evaluation, we follow

the approach by Zeller et al. (2013), who evaluated on a manually-annotated sample of lemma pairs. Working at the level of pairs makes manual annotation a simpler task: it is arguably much easier to decide for an individual pair of lemmas whether they are derivationally related, than to annotate a complete derivational family. This is in particularly true if one wishes to distinguish the cases where derivationally related words are also semantically related from the cases where the semantic relation is absent. However, there are some subtleties involved with pair-based evaluation, which we describe below.

5.1. Sample Construction

As noted by Zeller et al. (2013), random sampling of lemma pairs would make the evaluation unrealistic because most pairs are derivationally unrelated. To remedy this, Zeller et al. (2013) use two samples, one for measuring the precision and another for measuring the recall. The latter sample is obtained by sampling from the set of possibly derivationally related pairs, identified as such using a number of string similarity measures. More precisely, given lemma l_1 as input, they retrieve k lemmas that are most similar to l_1, and consider these as possibly derivationally related lemmas.

While the procedure of Zeller et al. (2013) is sound, it makes the interpretation of results somewhat difficult. Another problem is that the procedure overestimates the recall: the string-distance pooling strategy they use cannot guarantee that all related pairs will be retrieved. Some derivationally related pairs may be orthographically too different to fall within the top k most similar lemmas. On the other hand, retrieving too many lemmas will only increase the true negative rate and make the evaluation unrealistic.

We use a more straightforward methodology that remedies both problems identified above. First, we use only one sample on which we measure both precision and recall. Secondly, we employ an iterative sample construction method that ensures a reliable estimation of recall. The sampling procedure consists of two steps: (1) semi-automated construction of derivational families and (2) sampling of lemma pairs from the obtained derivational families.

Step 1: Construction of derivational families. We start with a set of lemmas, chosen randomly from the corpus. These lemmas constitute the initial singleton derivational families. Using string similarity measures, for each derivational family $\{l\}$, we retrieve from the inflectional lexicon $\mathcal{L}$ a set of lemmas $\{l_1, \ldots, l_n\}$ that are similar to l. The annotator inspects the retrieved sets and adds to the derivational family all lemmas that are derivationally related to l. In the next iteration, the newly added lemmas are used to retrieve new and previously unconsidered lemma candidates, which are then again inspected by the annotator. The procedure continues until no more lemmas can be added to the derivational families, indicating that the derivational families are complete.

For this procedure to work, we must ensure that we miss no derivationally related lemmas when retrieving the lemmas. In each iteration, we retrieve a pool of candidate lemmas comprised of k most similar lemmas according to six string similarity measures: measure D_4 given by (1), measure D_3 (also proposed by Majumder et al. (2007) as a slight variation of D_4), the reversed versions of measures D_3 and D_4 (to capture prefixal variation), and 2-gram and 3-gram Dice-based measures proposed by Adamson and Boreham (1974). We use $k = 10$ in our experiments. Notice that, because we build our sample iteratively, always expanding from the fringe of a derivational family, k need not be large. We believe that with $k = 10$, and with the string-similarity measures that we used, our annotations are almost complete with respect to recall.

Using the described procedure, we built a sample consisting of 481 lemmas from lexicon $\mathcal{L}$ grouped into 50 complete derivational families.

Step 2: Sampling of lemma pairs. In the second step, we sample lemma pairs from the derivational families. We sample 2000 within-family pairs (positive pairs) and 2000 pairs in which one lemma is not a member of the family (negative pairs). Although negative pairs could simply be sampled from pairs of lemmas not belonging to the same family, such pairs tend to be orthographically too different and would yield unrealistic precision estimates. Instead, for each lemma l_1 from each derivational family we retrieve a set of ten lemmas that are most similar to l_1 but do not belong to the same family as l_1, and then sample lemma pairs from this set. To measure string similarity, we use measure D_4 and its reversed version, to account for prefix derivations.

Finally, we split the resulting 4000 pairs into a test set and a development set, each containing 1000 positive and 1000 negative pairs.

5.2. Gold Standard Annotation

The above-described sample construction procedure provides us with a sample of evenly distributed derivationally related and derivationally unrelated lemma pairs. However, to allow for a more insightful analysis, we decided to follow the approach of Zeller et al. (2013) and introduce additional categories. We subcategorize the positive pairs into two categories: **R** (derivationally and semantically related) and **M** (only derivationally related), and the negative pairs into three categories: **N** (no relation), **C** (compositional relation), and **L** (invalid lemma), as shown in Table 3. Note that in this work we treat both **R** and **M** as positive classes, i.e., we consider a derivationally related pair of lemmas to be positive even if there is no semantic relation between the lemmas.

A single annotator with linguistic expertise manually annotated each of the 2000 pairs from both samples into one of the five categories. The annotation amounts to disambiguating between **R** and **M** cases for the positive pairs, and between **N**, **C**, and **L** cases for the negative pairs. For the positive pairs, we additionally labeled cases of prefixal derivation as either **Rp** or **Mp**. Table 4 shows an excerpt from the test sample. We make this sample freely available.[1]

The distribution of labels is shown in Table 5. A couple of errors introduced in the previous steps (construction of derivational families) were corrected during the annotation, increasing the number of positive pairs to 1009. The majority of positive pairs (85.7%) are both derivationally as well

Label	Description	Example
R	morphologically and semantically related	*izgubiti*$_V$ – *izgubljen*$_A$ (*loose – lost*)
M	morphologically but not semantically related	*brat*$_N$ – *bratić*$_N$ (*brother – cousin*)
N	no morphological relation	*konzumacija*$_N$ – *konzultirati*$_V$ (*consumption – to consult*)
C	no derivational but compositional relation	*ruka*$_N$ – *rukomet*$_N$ (*hand – handball*)
L	invalid lemma (mislemmatization, wrong POS, foreign words)	*razuvjeriti*$_V$ – *razultat*$_N$ (*undeceive – n/a*)

Table 3: Categories for lemma pair classification

Label	Lemma 1	Lemma 2
N	*denacionaliziran*$_A$	*regionaliziran*$_A$
N	*čarobnjački*$_A$	*čarda*$_N$
N	*neomiljen*$_A$	*neoboriv*$_A$
R	*interpret*$_N$	*interpreter*$_N$
Rp	*nacionalizirati*$_V$	*supranacionalan*$_A$
L	*rabin*$_N$	*rabinov*$_N$
N	*frapirati*$_V$	*terapirati*$_V$
R	*prosvjed*$_N$	*prosvjedovanje*$_N$
Rp	*prebrinuti*$_V$	*zabrinutost*$_N$
N	*neusuglašen*$_A$	*ovlašen*$_A$
N	*prosvjednički*$_A$	*prosvjećivati*$_V$
L	*neraspoloženje*$_N$	*rasploženje*$_N$
R	*čarobnjakov*$_A$	*čarolija*$_N$
M	*konzumeristički*$_A$	*konzumov*$_A$
N	*izazivački*$_A$	*pokazivački*$_A$
N	*poslužilac*$_N$	*brazilac*$_N$
M	*konzumizam*$_N$	*konzumov*$_A$
R	*briga*$_N$	*brinuti*$_V$

Table 4: Excerpt from the test sample of lemma pairs

	R	**M**	**N**	**C**	**L**	Total
All pairs	855	154	770	46	175	2,000
Suffixation only	424	28	770	46	175	1,443

Table 5: Distribution of labels in the test sample

as semantically related (**R**). Since in this work we do not consider prefixal derivation, we filtered out all pairs labeled as **Rp** or **Mp**, which amounts to 27.9% of the sample. The resulting subsample consists of 1,443 pairs, of which 31.3% are positive (**R** or **M**). In this subsample the ratio of positive pairs that are both derivationally and semantically related is 93.8%; this increase is expected because suffixal derivation is more meaning preserving than prefixal derivation.

6. Results

6.1. Quantitative Analysis

Table 6 presents the overall results for unsupervised (U) and knowledge-based (K) acquisition method, initial (Kv0) and revised (Kv1) version. We compare against two baselines: a prefix stemmer, which truncates each word to first p letters (using $p = 6$ maximizes the F1-score on our sample) and a rule-based stemmer by Ljubešić et al. (2007).[3] DERIVBASE.HR outperforms both baselines by a wide margin.

[3]Note that this is an inflectional stemmer and therefore not suitable for clustering derivationally related words. We nonetheless include it here for the sake of completeness.

Method	# clusters	P	R	F_1
DERIVBASE.HR (U)	37,999	76.0	75.4	75.7
DERIVBASE.HR (Kv0)	57,157	78.6	55.1	64.8
DERIVBASE.HR (Kv1)	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

Table 6: Performance on the test sample

	P	R		P	R
N-N	74.3	74.3	N-A	87.7	77.2
A-A	**88.9**	78.1	N-V	76.4	74.3
V-V	72.2	68.4	A-V	84.6	**84.6**

Table 7: Precision and recall across different part of speech

The knowledge-based version reaches 81.2% precision and 76.5% recall, and outperforms the unsupervised version in terms of precision (difference by 5.2 percentage points), while recall is comparable. The revised version of DERIVBASE.HR, which was induced with additional 30 derivational patterns, has a substantially higher recall (increase in 21.4 percentage points).

Table shows the precision and recall scores across different part-of-speech combinations for the knowledge-based DERIVBASE.HR. The recall is lowest for verb-verb pairs, suggesting that we still lack coverage for verb-to-verb derivational patterns. Interestingly, Zeller et al. (2013) report a similar finding for their German resource.

6.2. Discussion

Although our results are very encouraging, we see a number of possibilities for improvement, especially for the knowledge-base method. Recall could in principle be improved by further inspection of the false positives and addition of new patterns. This, however, would not solve the problem of missing lemmas that hinder the transitive closure. A possible workaround might be to extend the set of derivational patters so that it also includes the compositions of two or more derivational patterns.

Improvements in precision could be achieved by restricting to a set of more confident patterns, as proposed by Zeller et al. (2013), where confidences could be defined manually or estimated from an annotated sample. Performing clustering based on these confidences (instead of doing a transitive closure) could also improve the precision.

Perhaps the weakest point of our approach is the low quality of the morphological lexicon given as input to the clustering method. As described in Section 4.2., the quality of the

inflectional lexicon is 55.0% precision and 34.4% recall. Improving the quality of the inflectional lexicon could substantially improve the precision and recall of the clusters.

7. Conclusion

We have described the construction and evaluation of DERIVBASE.HR, a morphological resource that groups 100k lemmas into 56k derivational families. The knowledge-based version of DERIVBASE.HR attains good clustering quality of 78.8% F_1-score. As with similar resources for other languages, we expect DERIVBASE.HR to be useful for a number of natural language processing tasks. For future work, we will focus on improving DERIVBASE.HR along the above-discussed lines. We will also consider prefixation, as well as the splitting up of derivational families according to semantic relatedness.

8. Acknowledgments

This work was supported by the Croatian Science Foundation (project 02.03/162: "Derivational Semantic Models for Information Retrieval").

9. References

Adamson, G. W. and Boreham, J. (1974). The use of an association measure based on character structure to identify semantically related pairs of words and document titles. *Information Processing and Management*, 10(7/8):253–260.

Agić, Ž., Ljubešić, N., and Merkler, D. (2013). Lemmatization and morphosyntactic tagging of Croatian and Serbian. In *4th Biennial International Workshop on Balto-Slavic Natural Language Processing (BSNLP 2013)*.

Amigó, E., Gonzalo, J., Artiles, J., and Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486.

Baayen, H. R., Piepenbrock, R., and Gulikers, L. (1996). *The CELEX Lexical Database. Release 2. LDC96L14*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Bilgin, O., Çetinoğlu, O., and Oflazer, K. (2004). Morphosemantic relations in and across Wordnets. In *Proceedings of the Global WordNet Conference*, pages 60–66, Brno, Czech Republic.

Ćavar, D., Jazbec, I.-P., and Stojanov, T. (2009). Cromomorphological analysis for standard croatian and its synchronic and diachronic dialects and variants. *Finite-State Methods and Natural Language Processing. Frontiers in Artificial Intelligence and Applications*, 19:183–190.

Fellbaum, C., Osherson, A., and Clark, P. (2009). Putting semantics into WordNet's "morphosemantic" links. In *Proceedings of the Third Language and Technology Conference*, pages 350–358, Poznań, Poland.

Gaussier, E. (1999). Unsupervised learning of derivational morphology from inflectional lexicons. In *ACL'99 Workshop Proceedings on Unsupervised Learning in Natural Language Processing*, pages 24–30, College Park, Maryland, USA.

Green, R., Dorr, B. J., and Resnik, P. (2004). Inducing frame semantic verb classes from wordnet and ldoce. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 375–382, Barcelona, Spain.

Habash, N. and Dorr, B. (2003). A categorial variation database for English. In *Proceedings of the Anuual Meeting of the North American Association for Computational Linguistics*, pages 96–102, Edmonton, Canada.

Halácsy, P., Kornai, A., and Oravecz, C. (2007). HunPos: An open source trigram tagger. In *Proceedings of ACL 2007*, pages 209–212, Prague, Czech Republic.

Hammarström, H. and Borin, L. (2011). Unsupervised learning of morphology. *Computational Linguistics*, 37(2):309–350.

Hana, J. (2008). Knowledge- and labor-light morphological analysis. *Ohio State University Working Papers in Linguistics*, 58:52–84.

Ingason, A. K., Helgadóttir, S., Loftsson, H., and Rögnvaldsson, E. (2008). A mixed method lemmatization algorithm using a hierarchy of linguistic identities (HOLI). In *Proceedings of GoTAL*, pages 205–216.

Koeva, S., Krstev, C., and Vitas, D. (2008). Morpho-semantic relations in WordNet–a case study for two Slavic languages. In *Proceedings of the Fourth Global WordNet Conference*, pages 239–254.

Ljubešić, N. and Erjavec, T. (2011). hrWaC and slWac: Compiling web corpora for Croatian and Slovene. In *Proceedings of Text, Speech and Dialogue*, pages 395–402, Plzeň, Czech Republic.

Ljubešić, N., Boras, D., and Kubelka, O. (2007). Retrieving information in Croatian: Building a simple and efficient rule-based stemmer. In *Digital information and heritage*, pages 313–320, Zagreb. Odsjek za informacijske znanosti Filozofskog fakulteta u Zagrebu.

Luong, M.-T., Socher, R., and Manning, C. D. (2013). Better word representations with recursive neural networks for morphology. In *Proceedings of the Conference on Natural Language Learning*, pages 104–113, Sofia, Bulgaria.

Majumder, P., Mitra, M., Parui, S. K., Kole, G., Mitra, P., and Datta, K. (2007). YASS: Yet another suffix stripper. *ACM Transactions on Information Systems*, 25(4):18:1–18:20.

Oliver, A. (2003). Use of internet for augmenting coverage in a lexical acquisition system from raw corpora. In *Workshop on Information Extraction for Slavonic and Other Central and Eastern European Languages (IESL 2003), RANLP*.

Padó, S., Šnajder, J., and Zeller, B. (2013). Derivational smoothing for syntactic distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 731–735, Sofia, Bulgaria.

Pala, K. and Hlaváčková, D. (2007). Derivational relations in Czech WordNet. In *Proceedings of the ACL Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, pages 75–81.

Piasecki, M., Ramocki, R., and Maziarz, M. (2012). Recognition of Polish derivational relations based on supervised learning scheme. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 916–922, Istanbul, Turkey.

Sagot, B. (2005). Automatic acquisition of a Slovak lexicon from a raw corpus. *Lecture Notes in Computer Science*, 3658:156–163.

Shnarch, E., Goldberger, J., and Dagan, I. (2011). A probabilistic modeling framework for lexical entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 558–563, Portland, OR.

Šnajder, J. and Dalbelo Bašić, B. (2008). Higher-order functional representation of Croatian inflectional morphology. In *Proceedings of the 6th International Conference on Formal Approaches to South Slavic and Balkan Languages*, pages 121–130, Dubrovnik, Croatia.

Šnajder, J. and Dalbelo Bašić, B. (2009). String distance-based stemming of the highly inflected Croatian language. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2009)*, pages 411–415, Borovets, Bulgaria.

Šnajder, J. and Dalbelo Bašić, B. (2010). A computational model of Croatian derivational morphology. In *Proceedings of the 7th International Conference on Formal Approaches to South Slavic and Balkan Languages*, pages 109–118, Dubrovnik, Croatia.

Šnajder, J., Dalbelo Bašić, B., and Marko, T. (2008). Automatic acquisition of inflectional lexica for morphological normalisation. *Information Processing and Management*, 44(5):1720–1731.

Šnajder, J., Padó, S., and Agić, Ž. (2013). Building and evaluating a distributional memory for Croatian. In *51st Annual Meeting of the Association for Computational Linguistics*, pages 784–789.

Šnajder, J. (2010). *Morfološka normalizacija tekstova na hrvatskome jeziku za dubinsku analizu i pretraživanje informacija*. Ph.D. thesis, University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb.

Šojat, K. and Srebačić, M. (2014). Morphosemantic relations between verbs in croatian wordnet. In *Seventh Global WordNet Conference*.

Tadić, M. and Fulgosi, S. (2003). Building the croatian morphological lexicon. In *Proceedings of the 2003 EACL Workshop on Morphological Processing of Slavic Languages*, pages 41–46. Association for Computational Linguistics.

Šnajder, J. and Bašić, B. D. (2010). A computational model of Croatian derivational morphology. *FASSBL7*, page 109.

Zeller, B., Šnajder, J., and Padó, S. (2013). Derivbase: Inducing and evaluating a derivational morphology resource for german. In *51st Annual Meeting of the Association for Computational Linguistics*.