Towards Semantic Validation of a Derivational Lexicon

Britta D. Zeller* **Sebastian Padó†** **Jan Šnajder§**
*Heidelberg University, Institut für Computerlinguistik
69120 Heidelberg, Germany
†Stuttgart University, Institut für maschinelle Sprachverarbeitung
70569 Stuttgart, Germany
§University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
zeller@cl.uni-heidelberg.de pado@ims.uni-stuttgart.de jan.snajder@fer.hr

Abstract

Derivationally related lemmas like *friend*$_N$ – *friendly*$_A$ – *friendship*$_N$ are derived from a common stem. Frequently, their meanings are also systematically related. However, there are also many examples of derivationally related lemma pairs whose meanings differ substantially, e.g., *object*$_N$ – *objective*$_N$. Most broad-coverage derivational lexicons do not reflect this distinction, mixing up semantically related and unrelated word pairs.

In this paper, we investigate strategies to recover the above distinction by recognizing semantically related lemma pairs, a process we call *semantic validation*. We make two main contributions: First, we perform a detailed data analysis on the basis of a large German derivational lexicon. It reveals two promising sources of information (distributional semantics and structural information about derivational rules), but also systematic problems with these sources. Second, we develop a classification model for the task that reflects the noisy nature of the data. It achieves an improvement of 13.6% in precision and 5.8% in F1-score over a strong majority class baseline. Our experiments confirm that both information sources contribute to semantic validation, and that they are complementary enough that the best results are obtained from a combined model.

1 Introduction

Morphological processing forms the first step of virtually all linguistic processing toolchains in natural language processing (NLP) and precedes other analyses such as part of speech tagging, parsing, or named entity recognition. There are three major types of morphological processes: (a) *Inflection* modifies word forms according to the grammatical context; (b) *derivation* constructs new words from individual existing words, typically through affixation; (c) *composition* combines multiple words into new lexical items. Computational treatment of morphology is often restricted to normalization, such as *lemmatization* (covering inflection only) or *stemming* (covering inflection and derivation heuristically, Porter (1980)).

An important reason is that English is morphologically a relatively simple language. Composition is not marked morphologically (*zoo gate*) and an important derivational pattern is *zero derivation* where the input and output terms are identical surface forms (*a fish / to fish*). Thus, lemmatization or stemming go a long way towards treating the aspects of English morphology relevant for NLP. The situation is different for languages with a complex morphology that calls for explicit treatment. In fact, recent years have seen a growing body of computational work in particular on derivation, which is a very productive process of word formation in Slavic languages but also in languages more closely related to English, like German (Štekauer and Lieber, 2005).

Derivation comprises a large number of distinct patterns, many of which cross part of speech boundaries (nominalization, verbalization, adjectivization), but some of which do not (gender indicators like *master / mistress*, approximations like *red / reddish*). A simple way to conceptualize derivation is that it partitions a language's vocabulary into *derivational families* of derivationally related lemmas (cf. Zeller et al. (2013), Gaussier (1999)). In WordNet, this type of information has been included to some extent by so-called "morpho-semantic" relations (Fellbaum et al., 2009), and the approach has been applied to languages other

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1728–1739, Dublin, Ireland, August 23-29 2014.

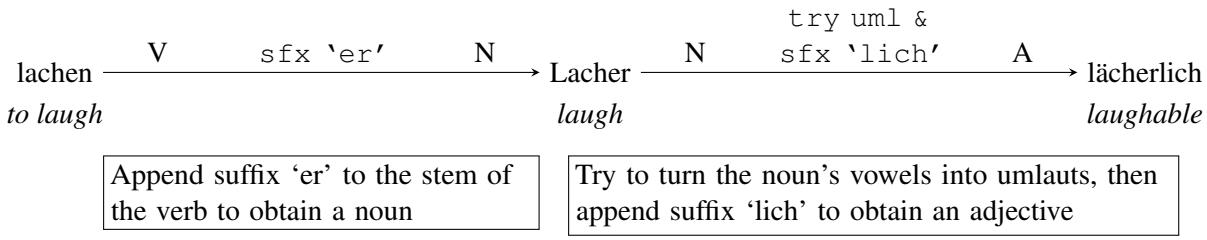

Figure 1: (Part of) a derivational family from DERIVBASE including derivational rules

than English (Bilgin et al., 2004; Pala and Hlaváčková, 2007). Another source of derivational information are stand-alone derivational lexicons such as CatVar (Habash and Dorr, 2003) for English, DERIVBASE (Zeller et al., 2013) for German, or the multilingual CELEX (Baayen et al., 1996).

Recent work has demonstrated that NLP can benefit from derivational knowledge. Shnarch et al. (2011) employ derivational knowledge in recognizing English textual entailment to better gauge the semantic similarity of text and hypothesis. Padó et al. (2013) improve the prediction of German semantic similarity judgments for lemma pairs by backing off to derivational families for infrequent lemmas. Luong et al. (2013) and Lazaridou et al. (2013) improve distributional semantic representations.

Note that all of these applications make use of derivational knowledge to address various *semantic* tasks, working on the assumption that derivationally related words, as represented in derivational lexicons, are strongly semantically related. This assumption is not completely warranted, though. The development of wide-coverage derivational lexicons is generally driven by morphological information, using for example finite-state technology (Karttunen and Beesley, 2005) to characterize known derivational patterns in terms of string transformations. Even though there is a strong correlation in derivation between morphology and semantics, it is not perfect. The absence of (synchronic) semantic relatedness can have a number of reasons, including accidental instantiation of derivational patterns (*corn – corner*) and diachronic meaning drift (*dog* (*animal*) – *dogged* (*determined*)). In other words, a substantial number of the lemma pairs in those lexicons are false positives regarding the level of semantic relatedness.

Our goal in this paper is to ameliorate this situation by developing strategies for the *semantic validation* of derivational lexicons, i.e., methods to determine, for lemma pairs that are derivationally related at the morphological level, whether they are in fact semantically related. We base our study on the German derivational lexicon DERIVBASE, and start by assessing which strategies can be used for its semantic validation (Section 2). In Sections 3 and 4, we analyze the contributions of semantic information (distributional semantics) as well as structural information (derivational rules). On the basis of our observations, we train a classifier that is able to semantically validate DERIVBASE at 89.9% F_1-score (Section 5), significantly outperforming a majority-class baseline of 84.1%. Section 6 reviews related work. Section 7 concludes the paper and outlines future work.

2 A Lexicon for German Derivation

2.1 DERIVBASE

DERIVBASE (Zeller et al., 2013) is a freely available derivational lexicon for German.[1] We used a rule-based framework to define derivation rules that cover suffixation, prefixation, and zero derivation as well as stem changes. Following the work of Šnajder and Dalbelo Bašić (2010), derivational processes are defined using derivational rules and higher-order string transformation functions. The only requirements for this method are (a) a comprehensive set of lemmas and (b) knowledge about admissible derivational rules, which can be gathered, for example, from linguistics textbooks.

Figure 1 shows a small sample from a derivational family with three lemmas and two derivational rules, one turning a verb into the corresponding event noun (in this case a semelfactive), and one turning the event into an adjective associated with it. Note that there are two perspectives on such a family: It can

[1] http://www.cl.uni-heidelberg.de/~zeller/res/derivbase/

DERIVBASE release	"Positive" class	Precision %	Recall %
1.2 (Zeller et al., 2013)[3]	R and M	83.0	71.0
1.4 (our analysis)	R and M	85.1	91.4
1.4 (our analysis)	R only	76.7	93.8

Table 1: DERIVBASE evaluation across releases on the DERIVBASE release 1.2 P and R samples

either be seen as a set of lemmas, or as a set of (independent) lemma pairs. We will assume the latter perspective in this paper, leaving questions of global coherence for future work.

DERIVBASE is a good example for the problems sketched in Section 1. It is defined purely on morphological grounds, without semantic validation of derivational families. Consequently, it contains a substantial number of words that are not semantically related.

2.2 Morphological and Semantic Relatedness in DERIVBASE

Our original evaluation of the quality of DERIVBASE in Zeller et al. (2013) was based on manually classified samples of lemma pairs. We introduce two samples, the "R sample", drawn from a large population of lemma pairs with high string similarity, in order to calculate recall, and the "P sample", drawn from the DERIVBASE families, in order to compute precision. Each lemma pair was classified into one of five categories (**R**: morphologically and semantically related; **M**: only morphologically related; **N**: not related; **L**: lemmatization errors; **C**: compounds) and inter-annotator agreement was checked to be substantial.[2] The overall best model (L123) showed 83% precision and 71% recall. However, this evaluation is limited in two important respects. First, it refers to DERIVBASE release 1.2 from 2013. Since then, we have extended DERIVBASE, e.g., with rules covering particle verbs, a very productive area of German derivation. Secondly, and more seriously, the previous evaluation considered all instances of **R** *and* **M** as true positives. In other words, in Zeller et al. (2013) we only evaluated the morphological relatedness of the lemma pairs but not the semantic relatedness.

We therefore start by presenting an evaluation of DERIVBASE focusing on the **R** instances in Table 1, reusing the DERIVBASE 1.2 "P" and "R" samples introduced in Zeller et al. (2013, see there for evaluation details). Between DERIVBASE 1.2 and 1.4, precision increased marginally and recall substantially, due mainly to the inclusion of rules that cover particle verbs. However, the numbers change substantially when only **R** (truly semantically related pairs) are counted as true positives. Recall increases by about 2.5%, but precision drops about 8.5%. Almost one quarter of all pairs in the lexicon are *not* semantically related.

A possible confounder of this analysis is that the "P sample" was drawn on DERIVBASE 1.2 and therefore does not include the novel items in DERIVBASE 1.4. We therefore created a novel DERIVBASE *1.4 extended sample* by combining the existing "P sample" with those pairs from the "R sample" that are in the coverage of a DERIVBASE rule as of DERIVBASE 1.4, resulting in 2,545 lemma pairs.

This DERIVBASE 1.4 extended sample will form the basis of all our analyses in this paper. The class distribution in the new sample is similar, but not identical, to the old P sample, as shown in Table 2. The relative frequency of **R** drops another 2%. Since this number also corresponds to the precision of the resource, the precision of the extended sample is 74.6%.

There are almost no compound errors **C**, which is not surprising given the rule-based construction of the lexicon, and only a relatively small number (about 5%) of lemmatization errors **L**, which fall outside the scope of our work. In contrast, both **N** and **M** occur with substantial frequency: Each class accounts for around 10% of the pairs. An analysis of **N** shows many cases of rule overgeneration: These are often pairs of lemmas whose stems are sufficiently similar that they might be related, e.g., by stem-changing derivation rules. Although such rules are valid in other contexts ($Verkauf_N$ – $Verkäufer_N$ (*selling – seller*)),

[2] Although we believe semantic relatedness to be fundamentally a graded scale, we adopt a binary notion of it as a convenient operational simplification that is supported by the good inter-annotator agreement for manual labeling in DERIVBASE.

[3] DERIVBASE 1.2 corresponds to DERIVBASE "L123" in (Zeller et al., 2013, p. 1207).

	R	M	N	L	C
Frequency	1899	265	240	131	8
Percentage overall	74.6	10.4	9.5	5.2	0.3
Percentage on dev. set	75.5	10.3	9.0	4.8	0.3
Percentage of test set	72.6	10.6	10.6	5.9	0.3

Table 2: Class distribution in our new DERIVBASE 1.4 extended sample

erroneous application leads to **N** cases like *Blase*$_N$ – *Bläser*$_N$ (*bubble – blower*). Also, we find false matches of common noun rules with named entities (*Empire*$_N$ – *Empirismus*$_N$ (*Empire – empiricism*)).

In contrast, many cases of **M** (as sketched in Section 1) refer to different senses of the same stem. As an example, consider *beruhen*$_V$ – *unruhig*$_A$ (*to rest on – restless*), both related to *Ruhe*$_N$ *(rest)*. In other cases, one of the two lemmas appears to have undergone a meaning shift (*Rappel*$_N$ – *rappeln*$_V$ (*craze – to rattle*)). This is particularly prominent for particle verbs (*bauen*$_V$ – *erbaulich*$_A$ (*build – edifying*)).

We divide the DERIVBASE 1.4 extended sample into a development and a test partition (70:30 ratio); the subsequent analyses consider only the development set.

2.3 Hypotheses for Semantic Validation

The preceding analysis of DERIVBASE has established that the lexicon contains a substantial number (around one fourth) of lemma pairs that are not semantically related. Therefore, it is in need of *semantic validation*, i.e., a computational procedure that can filter out semantically unrelated words.

In this paper, we frame semantic validation as a binary classification task that classifies all lemma pairs within one derivational family as either semantically related or unrelated. We consider this a first step towards splitting the current, morphologically motivated, DERIVBASE families into smaller, semantically coherent, families. We base our work on two general hypotheses about the types of information that might be helpful in this endeavor.

Hypothesis 1. *Distributional similarity indicates semantic relatedness between derivationally related words.* The instances of polysemy and meaning shift that we observe, in particular in the **M** class, motivate the use of distributional similarity (Turney and Pantel, 2010) since we expect these lemma pairs to be distributionally less related than cases of true semantic relatedness.

Hypothesis 2. *Derivational rules differ in their reliability.* Both the evidence from **M** and **N** indicate that some rules are more meaning-preserving than others. We expect this to be tied to both lexical properties of the rules (particle verbs are more likely than diminutives to radically change meaning) as well as structural properties (more specific rules are presumably more precise than generic rules).

In the two following Sections, we will operationalize these hypotheses and analyze the development set of the DERIVBASE 1.4 extended sample with respect to their empirical adequacy.

3 Analysis 1: Distributional Similarity for Semantic Validation

3.1 Measuring Distributional Similarity

We examine semantic similarities as predicted by simple bag-of-words semantic space models built from the lemmatized SDeWaC (Faaß et al., 2010), a large German web corpus containing about 880 million words. We compute vectors for all words covered in DERIVBASE using a window of $\pm$5 words within sentence boundaries and considering the 10k most frequent lemma-part of speech combinations of nouns, verbs, and adjectives in SDeWaC as contexts. Distributional vectors are built from co-occurrences which are measured with Local Mutual Information (Evert, 2005). The semantic similarity is measured by the cosine similarity between the vectors. Despite the size of the corpus, many lemmas from DERIVBASE occur very infrequently, and due to the inflection in German, it is important to retrieve as many occurrences of each lemma as possible.

We therefore use a very permissive two-step lemmatization scheme that starts from lemmas from the lexicon-based TreeTagger (Schmid, 1994), which provides reliable lemmas but with relatively low coverage, and supplements them with lemmas and parts of speech produced by the probabilistic MATE toolkit (Bohnet, 2010) when TreeTagger abstains.

3.2 Frequency Considerations

The advantage of the string transformation-based construction of DERIVBASE is its ability to include infrequent lemmas in the lexicon, and in fact DERIVBASE includes more than 250,000 content lemmas, some of which occur not more than three times in SDeWaC. However, this is a potential problem when we build distributional representations for all lemmas in DERIVBASE since it is known from the literature that similarity predictions for infrequent lemmas are often unreliable (Bullinaria and Levy, 2007).

Our data conform to expectations in this regard – infrequent lemmas are indeed problematic for validating the semantic relatedness of lemma pairs. More specifically, the semantic similarity of *related* lemmas (**R**) is systematically underestimated, because the lemma pairs from our sample are often too infrequent to share any dimensions. Consequently, they receive a low or zero cosine even when they are semantically strongly related. For example, each of the lemmas *Drogenverkauf*$_N$ – *Drogenverkäufer*$_N$ (*drug selling – drug seller*) has only nine lemmas as dimensions, and those are completely disjoint. This underestimation constitutes a general trend. The model assigns cosine scores below 0.1 to 64% of the related pairs in the development set, cosines below 0.2 to 81%, and cosines below 0.3 to 87%. Such low scores are problematic for separating related from unrelated pairs.

Two-step lemmatization is important for the proper handling of infrequent words. Compared to just using TreeTagger, the TreeTagger+MATE vectors for *auferstehen*$_V$ – *auferstehend*$_A$ (*to resurrect – resurrecting*) share seven more dimensions, including *Jesus*, *Lord*, *myth*, and *suffering*. Correspondingly, the cosine value of this pair rises by 50%. Generally, the amount of zero cosines in the DERIVBASE 1.4 extended sample drops by 45% using two-step lemmatization compared to one-step TreeTagger lemmatization.

3.3 Conceptual Considerations

In addition to the frequency considerations discussed above, we find three conceptual phenomena that affect distributional similarity independently of the frequency aspects.

The first one is the influence of *parts of speech*. Derivational rules often change the part of speech of the input lemma, and the parts of speech of its context words change as well. This decreases context overlap. For example, *Überschätzung*$_N$ – *überschätzt*$_A$ (*overestimate – overestimated*) is assigned a cosine of merely 0.09. The upper half of Table 3 shows the top ten individual and shared context words for this pair, ranked by LMI. The context words of the noun are mainly nominal heads of genitive complements (*overestimation of possibility/force/...*), while the context words of the adjective comprise many adverbs (*totally, widely, ...*). None of the shared contexts rank among of the top ten for both target lemmas. This is even more surprising considering that German adjectives and adverbs have the same surface realization (as opposed to English) and are more likely to form matching context words.

The second phenomenon that we identified as influencing semantic similarity is *markedness* (Battistella, 1996). A considerable number of derivational rules systematically produce marked terms. A striking example is the feminine suffix "*-in*" as in *Entertainer*$_N$ – *Entertainerin*$_N$: Although the lemmas are intuitively very similar, their cosine is as low as 0.1. The reason is that the female versions tend to be used in contexts where the gender of the entertainer is relevant. This is illustrated in the lower half of Table 3. The first two contexts for both words (*actor, singer*) stem from frequent enumerations (*actor and entertainer X*) and are almost identical, but again the female versions are marked for gender. We also find two female given names. As a result, the target lemmas receive a low distributional similarity.

The third example are cases of mild meaning shifts that were tagged by the annotators as **R**. These are lemmas where the semantic relatedness is intuitively clearly recognizable but may be accompanied by pretty substantial changes in the distribution of contexts. Consider the semantically related pair *Absteiger*$_N$ – *absteigend*$_A$ (*descender (person) – descending/decreasing*). It achieves only a cosine of

word pair (l_1, l_2)	context(l_1)	context(l_2)	shared contexts(l_1, l_2)
Überschätzung – überschätzt (*overestimation – overestimated*), $cos = 0.09$	eigen (*own*) warnen (*to alert*) Möglichkeit (*possibility*) führen (*to lead*) Kraft (*force*) Bedeutung (*meaning*) Fähigkeit (*ability*) Leistungsfähigkeit (*performance*) neigen (*to tend*) Einfluß (*influence*)	völlig (*totally*) Problem (*problem*) Gefahr (*danger*) Autor (*author*) weit (*widely*) total (*totally*) ernst (*seriously*) überhöht (*excessive*) gnadenlos (*mercilessly*) Hollywood (*Hollywood*)	völlig (*totally*) Möglichkeit (*possibility*) Bedeutung (*meaning*) Gefahr (*danger*) Einfluß (*influence*) überhöht (*excessive*) Macht (*power*) gnadenlos (*mercilessly*) Kraft (*force*) häufig (*frequent*)
Entertainer – Entertainerin (*entertainer – female entertainer*), $cos = 0.1$	Sänger (*singer*) Schauspieler (*actor*) Musiker (*musician*) Harald (*male given name*) Moderator (*anchorman*) Schmidt (*surname*) groß (*big*) Künstler (*artist*) Talent (*talent*) gut (*good*)	Sängerin (*female singer*) Schauspielerin (*actress*) Helga (*female given name*) Mutter (*mother*) berühmt (*famous*) brillant (*brilliant*) Lisa (*female given name*) Künstlerin (*female artist*) verstorben (*deceased*) Talent (*talent*)	Schauspieler (*actor*) Musiker (*musician*) **Talent** (*talent*) bekannt (*well-known*) Sängerin (*female singer*) beliebt (*popular*) groß (*big*) berühmt (*famous*) Sportler (*sportsman*) Schauspielerin (*actress*)

Table 3: Top ten individual and shared context words for *Überschätzung*$_N$ *– überschätzt*$_A$ (*overestimation – overestimated*) and *Entertainer*$_N$ *– Entertainerin*$_N$. Individual context words are ranked by LMI, shared context words by the product of their LMIs for the two target words. Shared context words that occur in the top ten contexts for both words are marked in **boldface**.

0.005, because *Absteiger* is almost exclusively used to refer to relegated sport teams while *absteigend* is used as a general verb of scalar change.

3.4 Ranking of Distributional Information

Given the results reported above, the standard distributional approach of using plain cosine scores to measure the absolute amount of co-occurrences does not seem very promising, due to the low absolute numbers of shared dimensions of the two lemmas. We expect other similarity measures, e.g., the Lin measure (Lin, 1998), to perform equally poorly since they do not change the fundamental approach. Also, although using a large corpus for semantic space construction might ameliorate the situation, we would prefer to make improvements on the modeling side of semantic validation.

We follow the ideas of Hare et al. (2009) and Lapesa and Evert (2013) who propose to consider semantic similarity in terms of ranks rather than absolute values. The advantage of rank-based similarity is that it takes the density of regions in the semantic space into account. That is, a low cosine value does not necessarily indicate low semantic relatedness – provided that the two words are located in a "sparse" region. Conversely, a high cosine value can be meaningless in a densely populated region. A second conceptual benefit of rank-based similarity is that it is directed: It is possible to distinguish the "forward" rank (the rank of l_1 in the neighborhood of l_2) and the "backward" rank (the rank of l_2 in the neighborhood of l_1). The previous studies found rank-based similarity to be beneficial for the prediction of priming results. In our case, it suggests a refined version of our Hypothesis 1:

Hypothesis 1'. *High* rank-based *distributional similarity indicates semantic relatedness between derivationally related words.*

4 Analysis 2: Derivational Rules for Semantic Validation

As discussed in Section 2.3, a second source of information that should be able to complement the problematic distributional similarity is provided by the derivational rules that are encoded in DERIVBASE (cf. the arrows in Figure 1). Our intuition is that some rules are "semantically stable", meaning that they reliably connect semantically similar lemmas, while other rules tend to cause semantic drifts. To examine

this situation, we perform a qualitative analysis on all lemma pairs connected by rule paths of length one ("simplex paths"), which are easy to analyze. Longer paths ("complex paths") are considered below.

We find that rules indeed behave differently. For example, the "*-in*" female marking rule from Section 3.3 is very reliable: every lemma pair connected by this rule is semantically related. At the other end of the scale, there are rules that consistently lead to semantically unrelated lemmas, e.g., the *"ver-"* noun-verb prefixation: *Zweifel*$_N$ – *verzweifeln*$_V$ (*doubt – to despair*). Foreign suffixes like "*-ktiv*" in *instruieren*$_V$ – *instruktiv*$_A$ (*to instruct – instructive*) retain semantic relatedness in most cases, but sometimes link actually unrelated lemmas (**N**, **C**, **L**). For example, *Objektiv*$_N$ – *Objektivismus*$_N$ (*lens – objectivism*), is an **N** pair for the suffix "*-ismus*". Finally, zero derivations and very short suffixes are less reliable: Since they easily match, they are often applied to incorrectly lemmatized words (**L**). For example, the "*-n*" suffix, which relates nationalities with countries (*Schwede*$_N$ – *Schweden*$_N$ (*Swede – Sweden*)). It matches many wrongly lemmatized nouns due to its syncretism with the plural dative/accusative suffix *-n*, as in *Schweineschnitzel*$_N$ – *Schweineschnitzeln*$_N$ (*pork cutlet – pork cutlets*$_{dat/acc\text{-}pl}$). This suggests that *rule-specific reliability* is a promising feature for semantic validation. Fortunately, due to its construction, DERIVBASE provides a rule chain for each lemma pair so that these reliabilities can be "read off". For other rules, however, the variance of the individual lemma pairs that instantiate the rule is large, and the applicability of the rule is influenced by the particular combination of rule and lemma pair. Such cases suggest that distributional knowledge and structural rule information should be combined, a direction that we will pursue in the next section.

On word pairs that are linked by "complex paths", i.e., more than one rule (*lachen*$_V$ – *lächerlich*$_A$ in Figure 1), our main observation in this respect is that rule paths show a clear "weakest link" property. One unreliable rule can be sufficient to cause a semantic drift, and only a sequence of reliable rules is likely to link two semantically related words. We will act on this observation in the next section.

5 A Machine Learning Model for Semantic Validation

5.1 Classification

The findings of our analyses suggest that the decision to classify lemma pairs as semantically related or unrelated can draw on a range of considerations. We therefore decided to adopt a machine learning approach and phrase semantic validation as a binary classification task, using the analyses we performed in Sections 3 and 4 as motivation for feature definition.

We train a classifier on the development portion of the DERIVBASE 1.4 extended sample (1,780 training instances, cf. Section 2.2). We learn a binary decision: Semantic relatedness (**R**) vs. non-semantic relatedness (**M**, **N**, **C**, **L**) within derivationally related pairs. For classification, we use a nonlinear model: Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel. Using the RBF kernel allows us to capture the non-linear dependencies between the features.[4] We rely on LIBSVM (Chang and Lin, 2011), a well-known SVM implementation. We optimize the C and γ hyperparameters of the SVM model using 3-fold cross-validation on the training data (i.e., the development portion of the extended sample).

5.2 Features

Our analyses motivate three feature groups comprising 35 individual features: Distributional, derivation rule-based ("structural"), and hybrid features. Table 4 gives a list.

Distributional features. All distributional features apply to the lemma or pair level. They are calculated from our BOW model with permissive lemmatization (Section 3.1). We use absolute and rank-based cosine similarity (Section 3.4) as well as the number of shared contexts (computed with LMI, cf. Section 3.3) and lemma frequency. To speed up processing, we compute the forward rank similarity for a lemma pair (l_1, l_2) not on the complete vocabulary but by pairing l_1 with a random sample of 1,000 lemmas from DERIVBASE (plus l_2 if it is not included). We do the computation analogously for the backward rank.

[4]The nonlinear SVM model outperforms a linear SVM. The difference is 0.8% F-Score, statistically significant at p=0.05.

Feature group (# features)	Type	Feature name (# features)	Description
Distributional (6)	l	Lemma frequency (2)	Normalized SDeWaC corpus lemma frequencies
	p	Cosine similarity	Standard cosine lemma similarity
	p	Dimensions shared	Number of shared context words
	p	Cos. rank similarity (2)	Rank-based forward and backward similarity
Structural (25)	r	Rule identity (11)	Indicator features for the top ten rules in the dev set + one aggregate feature for the rest
	r	Rule reliability	Percentage of rule applications on **R** pairs among all applications of the rule in dev set
	r	Rule frequency rank (2)	Rank-based rule frequency in DERIVBASE
	r	Avg. string distance (2)	Avg. Levenshtein distance for all rule instances
	p	POS combinations (6)	Indicator features for lemma POS combinations
	p	Path length	Length of the shortest path between the lemmas
	p	String distance (2)	Dice bigram coefficient; Levenshtein distance
Hybrid (4)	r	Average rank sim (2)	Frequency-weighted average rank similarity of rules on shortest path
	p	Rank sim deviation (2)	Difference between lemma pair rank similarity and average rule rank similarity

Table 4: Features used to characterize derivationally related lemma pairs. "Type" indicates the level at which each feature applies: *l* lemma level, *p* pair level, *r* rule level.

Structural features. The structural features encode properties of the rules and paths in DERIVBASE. Most features apply to the level of derivation rules. This includes the identity of the rule; its reliability (estimated as the ratio of its application on **R** pairs among all its applications on the dev set); its frequency rank among all rules (as a measure of specificity)[5]; and the average Levenshtein distance between the input and output lemmas (estimating rule complexity by measuring the amount of string modification).

For lemma pairs linked by complex paths (i.e., more than one rule, cf. Figure 1), the question arises how the rule-level features should be computed. Following our observations on "weakest link" behavior in Section 4, we always combine the feature values for the individual rules adopting the most pessimistic combination function (e.g., minimum in the case of reliability, maximum in the case of frequency rank).

Three more structural features are computed directly at the lemma pair level: their part of speech combination (e.g., *"NV"* for $oxide_N - oxidate_V$), the length of the shortest path connecting them, and the Levenshtein and Dice string distances between the two lemmas.

Hybrid features. Hybrid features combine rule-based and distributional information to avoid their respective shortcomings. We work with two hybrid features, one at rule level and one at pair level. The rule-level feature models the reliability of the rule. It is the average rank similarity for each rule (computed as a log frequency-weighted average over rule instances). This feature is a counterpart to rule reliability that is unsupervised in that it does not require class labels. We compute it by randomly drawing 200 lemma pairs for each rule from DERIVBASE (less if the rule has fewer instances). The pair-level feature is the difference between the rule's average rank similarity and the rank similarity for the current pair. It measures the rank of a pair relative to the rule's "baseline" rank and indicates how similar and dissimilar lemma pairs are compared to the rule average. In parallel to the structural features, values for complex rule paths are computed by minimum. Since the rank similarity is directional, we compute both hybrid features in two variants, one for each direction.[6]

[5]We compute this feature once only on simplex paths and once on all instances of the rule in DERIVBASE, trading reliability against noise.

[6]We also tested hybrid features based on raw cosine; however, this yielded worse results than the rank-based hybrid features.

Validation method	Precision	Recall	F_1	Accuracy
Majority baseline	72.6	**100**	84.1	72.6
Classifier, *only "cosine similarity" feature*	72.6	100	84.1	72.6
Classifier *only "similarity rank" feature*	80.3	90.3	85.0	76.8
Classifier, *only "rule identity" feature*	73.7	99.5	84.6	73.8
Classifier, *hybrid group*	80.4	95.3	87.2	79.7
Classifier, *distributional group*	80.5	96.6	87.8	80.5
Classifier, *structural group*	82.7	93.1	87.6	80.9
Classifier, *hybrid + distributional groups*	82.6	93.3	87.6	80.9
Classifier, *hybrid + structural groups*	84.9	93.7	89.1	83.4
Classifier, *distributional + structural groups*	85.3	94.6	89.7	84.3
Classifier, *all features*	**86.2**	93.9	**89.9**	**84.7**

Table 5: Accuracy, precision, recall, and F_1 on the test portion of the DERIVBASE 1.4 extended sample.

5.3 Results and Discussion

We applied the trained classifier to the test portion of the DERIVBASE 1.4 extended sample (cf. Section 2.2). Table 5 summarizes precision, recall, and F_1-score of the classifier for various combinations of features and feature groups. Recall that since our motivation is semantic validation, i.e., the removal of false positives, we are in particular interested in improving the *precision* of our predictions. We test significance of F_1 differences among models with bootstrap resampling (Efron and Tibshirani, 1993).

Our baseline is the majority class in the sample, **R**. Due to the sample's skewed class distribution (cf. Table 2), the frequency baseline is quite high (precision 72.6, F_1-score 84.1). We next consider the three most prominent individual features: Distributional similarity measured as cosine, distributional similarity measured as similarity rank, and rule identity. As expected from our analyses, the cosine similarity on its own is not reliable; in fact, it performs at baseline level. The rank-based similarity already leads to a considerable gain (precision +7.7%), but only a slight F_1-score increase of 0.9% that is not statistically significant at p=0.05. These results provide good empirical evidence for Hypothesis 1' (Section 3.4) and underscore that 1' is a more accurate statement than Hypothesis 1 (Section 2.3). On the structural side, rule identity alone improves the precision by 1.1%, with an F_1-score increase in 0.5% (again not significant).

We now proceed to complete feature groups, all of which perform at least 3% F_1-score better than the baseline, proving that the features within these groups are complementary. The hybrid feature group is the worst among the three. The distributional feature group is able to improve only slightly over the individual rank-based similarity feature in precision (80.5 vs. 80.3), but gains 6.3% in recall. This is sufficient for a significant improvement in F_1 (+3.7%, significant at p=0.01). The structural feature group performs surprisingly well, given that these features are very simple and most are computed only on the relatively small training set. It yields by far the highest precision (82.7), and its F_1-score is only slightly lower than the one of the distributional group (87.6 vs. 87.8). We take this as further evidence for the usefulness of structural information, as expressed by Hypothesis 2 (cf. Section 2.3).

Ultimately, all three feature groups turn out to be complementary. We obtain an improvement in F_1-score for two out of the three feature group combinations, and a clear improvement in precision in all cases. Finally, the best overall result is shown by the combination of all three feature groups. It attains an F_1-score of 89.9, an improvement of 5.8% over the baseline and 2.1% over the best feature group (both differences significant at p=0.01). Crucially, this model gains over 13% in precision while losing only 6% of recall compared to the baseline. This corresponds to a reduction of false positives in the sample by about half (from 27% to 14%) while the true positives were reduced only by 5% (from 73% to 68%).

Table 6 shows a breakdown of the predictions by the best model in terms of the five gold standard classes

	R	M	N	L	C	total
Gold annotation	554	81	81	45	2	763
Classified as **R**	520	36	16	29	2	603
Classified as **not R**	34	45	65	16	0	160

Table 6: Predictions on the test set of the *all features* Classifier per annotation class.

(**R**, **M**, **N**, **L**, **C**). Ignoring compounds (**C**), of which there are too few cases to analyze, we first find that the classifier achieves a high **R** recall. It is also very good in filtering out unrelated cases (**N**), of which it discards around 80%. The model recognizes morphologically but not semantically related word pairs (**M**) fairly well and manages to remove more than half of these. It has the hardest time with lemmatization errors (**L**), of which only about 35% were removed. However, this is not surprising: Lemmatization errors do not form a coherent category that would be easy to retrieve with the kinds of features that we have developed. We believe that such errors should be handled in an earlier stage, i.e., during preprocessing.

6 Related Work

Given that many derivational lexicons were only developed in recent years, we are only aware of one study (Jacquemin, 2010) that semantically validates the output of an existing derivational lexicon (Gaussier, 1999) to apply it to Question Answering. In contrast to our study, it requires elaborate dictionary information to look up which derivations are permitted for a specific lemma, as well as word sense disambiguation to determine the meaning of ambiguous words in context. Other related work comes from two areas: unsupervised morphology induction and semantic clustering.

Unsupervised morphology induction is concerned with the automatic identification of morphological relations (cf. Hammarström and Borin (2011) for an overview). Most approaches in this area do not differentiate between the inflectional and derivational level of morphology (Gaussier (1999) is an exception) and restrict themselves to the string level. Only a small number of studies (Schone and Jurafsky, 2000; Baroni et al., 2002) take distributional information into account.

Semantic clustering is the task of inducing semantic classes from (broadly speaking) distributional information (Turney and Pantel, 2010; im Walde, 2006). Boleda et al. (2012) include derivational properties in their feature set to learn Catalan adjective classes. However, the input to such studies is almost always a set of words from the same part of speech with no prior morphological constraints, while our input lemmas are morphologically preselected (via derivational rules), are often extremely infrequent, and exhibit systematical variation in parts of speech. To our knowledge, this challenging situation has not been addressed in previous studies.

Recent work has also considered the opposite problem, namely using derivational morphology for improving distributional similarity predictions. Luong et al. (2013) use recursive neural networks to learn representations of morphologically complex words and demonstrate the usefulness of their approach on word similarity tasks across different datasets. Similarly, Lazaridou et al. (2013) improve the word representations of derivationally related words by composing vector space representations of stems and derivational suffixes.

7 Conclusions

Almost all existing derivational lexicons do not distinguish between only morphologically related words on one hand and words that are both morphologically and semantically related words on the other hand. In this paper, we have addressed the task of recovering this distinction and called it *semantic validation*. We have used DERIVBASE, a German derivation lexicon, as the basis of our investigation.

We have made two contributions: (a) providing a detailed analysis of the types of information available for this task (distributional similarity as well as structural information about derivation rules) and the problems associated with each information type; and (b) training a machine learning classifier on linguistically

motivated features. The classifier, although not perfect, can substantially improve the precision of the word pairs in DERIVBASE and thus help to filter the derivational families in the lexicon. We are making this semantic validation information available in the DERIVBASE lexicon by attaching a probability for the class **R** to each lemma pair (see footnote 1 for the DERIVBASE URL).

The approach that we have described should transfer straightforwardly to other derivational lexicons and other languages on the conceptual level. The practical requirements are an appropriate corpus (for the distributional features) and derivational rule information (for the structural features).

There are two clear directions for future work. First, we plan to broaden our attention from word pairs to clusters and use the relatedness probabilities to cluster the derivational families in DERIVBASE into semantically coherent subfamilies. Second, we will demonstrate the impact of semantic validation on applications of derivational knowledge such as derivation-driven smoothing of distributional models (Padó et al., 2013).

Acknowledgments. We gratefully acknowledge partial funding by the European Commission (project EXCITEMENT (FP7 ICT-287923), first and second authors) as well as the Croatian Science Foundation (project 02.03/162: "Derivational Semantic Models for Information Retrieval", third author). We thank the reviewers for their valuable feedback.

References

Harald R. Baayen, Richard Piepenbrock, and Leon Gulikers. 1996. *The CELEX Lexical Database. Release 2. LDC96L14*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, Pennsylvania.

Marco Baroni, Johannes Matiasek, and Harald Trost. 2002. Unsupervised Discovery of Morphologically Related Words Based on Orthographic and Semantic Similarity. *Computing Research Repository*, cs.CL/0205006.

Edwin L. Battistella. 1996. *The Logic of Markedness*. Oxford University Press.

Orhan Bilgin, Ozlem Çetinoğlu, and Kemal Oflazer. 2004. Morphosemantic relations in and across Wordnets. In *Proceedings of the Global Wordnet Conference*, pages 60–66, Brno, Czech Republic.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the International Conference on Computational Linguistics*, pages 89–97, Beijing, China.

Gemma Boleda, Sabine Schulte im Walde, and Toni Badia. 2012. Modeling Regular Polysemy: A Study on the Semantic Classification of Catalan Adjectives. *Computational Linguistics*, 38(3):575–616.

John A. Bullinaria and Joe P. Levy. 2007. Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study. *Behavior Research Methods*, 39(3):510–526.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems Technology*, 2(3):27:1–27:27.

Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman and Hall, New York.

Stefan Evert. 2005. *The Statistics of Word Cooccurrences Word Pairs and Collocations*. Ph.D. thesis, University of Stuttgart.

Gertrud Faaß, Ulrich Heid, and Helmut Schmid. 2010. Design and Application of a Gold Standard for Morphological Analysis: SMOR in Validation. In *Proceedings of the Conference on Language Resources and Evaluation*, pages 803–810, Valletta, Malta.

Christiane Fellbaum, Anne Osherson, and Peter Clark. 2009. Putting semantics into WordNet's "morphosemantic" links. In *Proceedings of Human Language Technology. Challenges of the Information Society*, pages 350–358, Poznań, Poland.

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

Nizar Habash and Bonnie Dorr. 2003. A categorial variation database for English. In *Proceedings of the North American Association for Computational Linguistics*, pages 96–102, Edmonton, Canada.

Harald Hammarström and Lars Borin. 2011. Unsupervised Learning of Morphology. *Computational Linguistics*, 37(2):309–350.

Mary Hare, Michael Jones, Caroline Thomson, Sarah Kelly, and Ken McRae. 2009. Activating Event Knowledge. *Cognition*, 111(2):151–167.

Sabine Schulte im Walde. 2006. Experiments on the Automatic Induction of German Semantic Verb Classes. *Computational Linguistics*, 32(2):159–194.

Bernard Jacquemin. 2010. A derivational rephrasing experiment for question answering. In *Proceedings of the Conference on Language Resources and Evaluation*, pages 2380–2387, Valletta, Malta.

Lauri Karttunen and Kenneth R. Beesley. 2005. Twenty-five Years of Finite-state Morphology. In *Inquiries into Words, Constraints and Contexts. Festschrift for Kimmo Koskenniemi on his 60th Birthday*, pages 71–83. CSLI Publications, Stanford, California.

Gabriella Lapesa and Stefan Evert. 2013. Evaluating neighbor rank and distance measures as predictors of semantic priming. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 66–74, Sofia, Bulgaria.

Angeliki Lazaridou, Marco Marelli, Roberto Zamparelli, and Marco Baroni. 2013. Compositional-ly derived representations of morphologically complex words in distributional semantics. In *Proceedings of the Association for Computational Linguistics*, pages 1517–1526, Sofia, Bulgaria.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the International Conference on Machine Learning*, ICML, pages 296–304, San Francisco, California.

Minh-Thang Luong, Richard Socher, and Christopher D. Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Conference on Natural Language Learning*, pages 104–113, Sofia, Bulgaria.

Sebastian Padó, Jan Šnajder, and Britta Zeller. 2013. Derivational smoothing for syntactic distributional semantics. In *Proceedings of the Association for Computational Linguistics*, pages 731–735, Sofia, Bulgaria.

Karel Pala and Dana Hlaváčková. 2007. Derivational relations in Czech wordnet. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, pages 75–81, Prague, Czech Republic.

Martin Porter. 1980. An Algorithm for Suffix Stripping. *Program*, 14(3):130–137.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on Natural Language Processing*, Manchester, UK.

Patrick Schone and Daniel Jurafsky. 2000. Knowledge-free induction of morphology using latent semantic analysis. In *Proceedings of the Conference on Natural Language Learning*, pages 67–72. Lisbon, Portugal.

Eyal Shnarch, Jacob Goldberger, and Ido Dagan. 2011. A probabilistic modeling framework for lexical entailment. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 558–563, Portland, Oregon.

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

Pavol Štekauer and Rochelle Lieber, editors. 2005. *Handbook of Word-Formation*, volume 64 of *Studies in Natural Language and Linguistic Theory*. Springer.

Peter D. Turney and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

Britta Zeller, Jan Šnajder, and Sebastian Padó. 2013. DErivBase: Inducing and evaluating a derivational morphology resource for German. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1201–1211, Sofia, Bulgaria.