Compositional-ly Derived Representations of Morphologically Complex Words in Distributional Semantics

Angeliki Lazaridou and **Marco Marelli** and **Roberto Zamparelli** and **Marco Baroni**
Center for Mind/Brain Sciences (University of Trento, Italy)
first.last@unitn.it

Abstract

Speakers of a language can construct an unlimited number of new words through morphological derivation. This is a major cause of data sparseness for corpus-based approaches to lexical semantics, such as distributional semantic models of word meaning. We adapt compositional methods originally developed for phrases to the task of deriving the distributional meaning of morphologically complex words from their parts. Semantic representations constructed in this way beat a strong baseline and can be of higher quality than representations directly constructed from corpus data. Our results constitute a novel evaluation of the proposed composition methods, in which the *full additive* model achieves the best performance, and demonstrate the usefulness of a compositional morphology component in distributional semantics.

1 Introduction

Effective ways to represent word meaning are needed in many branches of natural language processing. In the last decades, corpus-based methods have achieved some degree of success in modeling lexical semantics. Distributional semantic models (DSMs) in particular represent the meaning of a word by a vector, the dimensions of which encode corpus-extracted co-occurrence statistics, under the assumption that words that are semantically similar will occur in similar contexts (Turney and Pantel, 2010). Reliable distributional vectors can only be extracted for words that occur in many contexts in the corpus. Not surprisingly, there is a strong correlation between word frequency and vector quality (Bullinaria and Levy, 2007), and since most words occur only once even in very large corpora (Baroni, 2009), DSMs suffer data sparseness.

While word rarity has many sources, one of the most common and systematic ones is the high *productivity* of morphological derivation processes, whereby an unlimited number of new words can be constructed by adding affixes to existing stems (Baayen, 2005; Bauer, 2001; Plag, 1999).[1] For example, in the multi-billion-word corpus we introduce below, perfectly reasonable derived forms such as *lexicalizable* or *affixless* never occur. Even without considering the theoretically infinite number of possible derived nonce words, and restricting ourselves instead to words that are already listed in dictionaries, complex forms cover a high portion of the lexicon. For example, morphologically complex forms account for 55% of the lemmas in the CELEX English database (see Section 4.1 below). In most of these cases (80% according to our corpus) the stem is more frequent than the complex form (e.g., the stem *build* occurs 15 times more often than the derived form *rebuild*, and the latter is certainly not an unusual derived form).

DSMs ignore derivational morphology altogether. Consequently, they cannot provide meaning representations for new derived forms, nor can they harness the systematic relation existing between stems and derivations (any English speaker can infer that to *rebuild* is to *build again*, whether they are familiar with the prefixed form or not) in order to mitigate derived-form sparseness problems. A simple way to handle derivational mor-

[1] Morphological *derivation* constructs new words (in the sense of lemmas) from existing lexical items (*resource+ful→resourceful*). In this work, we do not treat *inflectional* morphology, pertaining to affixes that encode grammatical features such as number or tense (*dog+s*). We use *morpheme* for any component of a word (*resource* and *-ful* are both morphemes). We use *stem* for the lexical item that constitutes the base of derivation (*resource*) and *affix* (*prefix* or *suffix*) for the element attached to the stem to derive the new form (*-ful*). In English, stems are typically independent words, affixes *bound* morphemes, i.e., they cannot stand alone. Note that a stem can in turn be morphologically derived, e.g., *point+less* in *pointless+ly*. Finally, we use *morphologically complex* as synonymous with *derived*.

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1517–1526,
Sofia, Bulgaria, August 4-9 2013.

phology would be to identify the stem of rare derived words and use its distributional vector as a proxy to derived-form meaning.[2] The meaning of *rebuild* is not that far from that of *build*, so the latter might provide a reasonable surrogate. Still, something is clearly lost (if the author of a text felt the need to use the derived form, the stem was not fully appropriate), and sometimes the jump in meaning can be quite dramatic (*resourceless* and *resource* mean very different things!).

In the past few years there has been much interest in how DSMs can scale up to represent the meaning of larger chunks of text such as phrases or even sentences. Trying to represent the meaning of arbitrarily long constructions by directly collecting co-occurrence statistics is obviously ineffective and thus methods have been developed to derive the meaning of larger constructions as a function of the meaning of their constituents (Baroni and Zamparelli, 2010; Coecke et al., 2010; Mitchell and Lapata, 2008; Mitchell and Lapata, 2010; Socher et al., 2012). *Compositional* distributional semantic models (cDSMs) of word units aim at handling, compositionally, the high productivity of phrases and consequent data sparseness. It is natural to hypothesize that the same methods can be applied to morphology to derive the meaning of complex words from the meaning of their parts: For example, instead of harvesting a *rebuild* vector directly from the corpus, the latter could be constructed from the distributional representations of *re-* and *build*. Besides alleviating data sparseness problems, a system of this sort, that automatically induces the semantic contents of morphological processes, would also be of tremendous theoretical interest, given that the semantics of derivation is a central and challenging topic in linguistic morphology (Dowty, 1979; Lieber, 2004).

In this paper, we explore, for the first time (except for the proof-of-concept study in Guevara (2009)), the application of cDSMs to derivational morphology. We adapt a number of composition methods from the literature to the morphological setting, and we show that some of these methods can provide better distributional representations of derived forms than either those directly harvested from a large corpus, or those obtained by using the stem as a proxy to derived-form meaning. Our results suggest that exploiting morphology could improve the quality of DSMs in general, extend the range of tasks that cDSMs can successfully model and support the development of new ways to test their performance.

[2]Of course, spotting and segmenting complex words is a big research topic unto itself (Beesley and Karttunen, 2000; Black et al., 1991; Sproat, 1992), and one we completely sidestep here.

2 Related work

Morphological induction systems use corpus-based methods to decide if two words are morphologically related and/or to segment words into morphemes (Dreyer and Eisner, 2011; Goldsmith, 2001; Goldwater and McClosky, 2005; Goldwater, 2006; Naradowsky and Goldwater, 2009; Wicentowski, 2004). Morphological induction has recently received considerable attention since morphological analysis can mitigate data sparseness in domains such as parsing and machine translation (Goldberg and Tsarfaty, 2008; Lee, 2004). Among the cues that have been exploited there is distributional similarity among morphologically related words (Schone and Jurafsky, 2000; Yarowsky and Wicentowski, 2000). Our work, however, differs substantially from this track of research. We do not aim at segmenting morphological complex words or identifying paradigms. Our goal is to automatically construct, given distributional representations of stems and affixes, semantic representations for the derived words containing those stems and affixes. A morphological induction system, given *rebuild*, will segment it into *re-* and *build* (possibly using distributional similarity between the words as a cue). Our system, given *re-* and *build*, predicts the (distributional semantic) meaning of *rebuild*.

Another emerging line of research uses distributional semantics to model human intuitions about the semantic transparency of morphologically derived or compound expressions and how these impact various lexical processing tasks (Kuperman, 2009; Wang et al., 2012). Although these works exploit vectors representing complex forms, they do not attempt to generate them compositionally.

The only similar study we are aware of is that of Guevara (2009). Guevara found a systematic geometric relation between corpus-based vectors of derived forms sharing an affix and their stems, and used this finding to motivate the composition method we term *lexfunc* below. However, unlike us, he did not test alternative models, and he only presented a qualitative analysis of the trajectories triggered by composition with various affixes.

3 Composition methods

Distributional semantic models (DSMs), also known as vector-space models, semantic spaces, or by the names of famous incarnations such as Latent Semantic Analysis or Topic Models, approximate the meaning of words with vectors that record their patterns of co-occurrence with corpus context features (often, other words). There is an extensive literature on how to develop such models and on their evaluation. Recent surveys include Clark (2012), Erk (2012) and Turney and Pantel (2010). We focus here on compositional DSMs (cDSMs). Since the very inception of distributional semantics, there have been attempts to compose meanings for sentences and larger passages (Landauer and Dumais, 1997), but interest in compositional DSMs has skyrocketed in the last few years, particularly since the influential work of Mitchell and Lapata (2008; 2009; 2010). For the current study, we have reimplemented and adapted to the morphological setting all cDSMs we are aware of, excluding the tensor-product-based models that Mitchell and Lapata (2010) have shown to be empirically disappointing and the models of Socher and colleagues (Socher et al., 2011; Socher et al., 2012), that require complex optimization procedures whose adaptation to morphology we leave to future work.

Mitchell and Lapata proposed a set of simple and effective models in which the composed vectors are obtained through component-wise operations on the constituent vectors. Given input vectors $\mathbf{u}$ and $\mathbf{v}$, the multiplicative model (**mult**) returns a composed vector $\mathbf{c}$ with: $c_i = u_i v_i$. In the weighted additive model (**wadd**), the composed vector is a weighted sum of the two input vectors: $\mathbf{c} = \alpha\mathbf{u} + \beta\mathbf{v}$, where α and β are two scalars. In the **dilation** model, the output vector is obtained by first decomposing one of the input vectors, say $\mathbf{v}$, into a vector parallel to $\mathbf{u}$ and an orthogonal vector. Following this, the parallel vector is dilated by a factor λ before re-combining. This results in: $\mathbf{c} = (\lambda - 1)\langle\mathbf{u}, \mathbf{v}\rangle\mathbf{u} + \langle\mathbf{u}, \mathbf{u}\rangle\mathbf{v}$.

Guevara (2010) and Zanzotto et al. (2010) propose the full additive model (**fulladd**), where the two vectors to be added are pre-multiplied by weight matrices: $\mathbf{c} = \mathbf{A}\mathbf{u} + \mathbf{B}\mathbf{v}$

Since the Mitchell and Lapata and *fulladd* models were developed for phrase composition, the two input vectors were taken to be, very straightforwardly, the vectors of the two words to be composed into the phrase of interest. In morphological derivation, at least one of the items to be composed (the affix) is a bound morpheme. In our adaptation of these composition models, we build bound morpheme vectors by accumulating the contexts in which a set of derived words containing the relevant morphemes occur, e.g., the *re-* vector aggregates co-occurrences of *redo*, *remake*, *retry*, etc.

Baroni and Zamparelli (2010) and Coecke et al. (2010) take inspiration from formal semantics to characterize composition in terms of function application, where the distributional representation of one element in a composition (the *functor*) is not a vector but a function. Given that linear functions can be expressed by matrices and their application by matrix-by-vector multiplication, in this *lexical function* (**lexfunc**) model, the functor is represented by a matrix $\mathbf{U}$ to be multiplied with the argument vector $\mathbf{v}$: $\mathbf{c} = \mathbf{U}\mathbf{v}$. In the case of morphology, it is natural to treat bound affixes as functions over stems, since affixes encode the systematic semantic patterns we intend to capture. Unlike the other composition methods, *lexfunc* does not require the construction of distributional vectors for affixes. A matrix representation for every affix is instead induced directly from examples of stems and the corresponding derived forms, in line with the intuition that every affix corresponds to a different pattern of change of the stem meaning.

Finally, as already discussed in the Introduction, performing no composition at all but using the stem vector as a surrogate of the derived form is a reasonable strategy. We saw that morphologically derived words tend to appear less frequently than their stems, and in many cases the meanings are close. Consequently, we expect a **stem**-only "composition" method to be a strong baseline in the morphological setting.

4 Experimental setup

4.1 Morphological data

We obtained a list of stem/derived-form pairs from the CELEX English Lexical Database, a widely used 100K-lemma lexicon containing, among other things, information about the derivational structure of words (Baayen et al., 1995). For each derivational affix present in CELEX, we extracted from the database the full list of stem/derived pairs matching its most common part-of-speech signature (e.g., for *-er* we only considered pairs

Affix	Stem/Der. POS	Training Items	HQ/Tot. Test Items	Avg. SDR
-able	verb/adj	177	30/50	5.96
-al	noun/adj	245	41/50	5.88
-er	verb/noun	824	33/50	5.51
-ful	noun/adj	53	42/50	6.11
-ic	noun/adj	280	43/50	5.99
-ion	verb/noun	637	38/50	6.22
-ist	noun/noun	244	38/50	6.16
-ity	adj/noun	372	33/50	6.19
-ize	noun/verb	105	40/50	5.96
-less	noun/adj	122	35/50	3.72
-ly	adj/adv	1847	20/50	6.33
-ment	verb/noun	165	38/50	6.06
-ness	adj/noun	602	33/50	6.29
-ous	noun/adj	157	35/50	5.94
-y	noun/adj	404	27/50	5.25
in-	adj/adj	101	34/50	3.39
re-	verb/verb	86	27/50	5.28
un-	adj/adj	128	36/50	3.23
tot	*/*	6549	623/900	5.52

Table 1: Derivational morphology dataset

having a verbal stem and nominal derived form). Since CELEX was populated by semi-automated morphological analysis, it includes forms that are probably not synchronically related to their stems, such as *crypt+ic* or *re+form*. However, we did not manually intervene on the pairs, since we are interested in training and testing our methods in realistic, noisy conditions. In particular, the need to pre-process corpora to determine which forms are "opaque", and should thus be bypassed by our systems, would greatly reduce their usefulness. Pairs in which either word occurred less than 20 times in our source corpus (described in Section 4.2 below) were filtered out and, in our final dataset, we only considered the 18 affixes (3 prefixes and 15 suffixes) with at least 100 pairs meeting this condition. We randomly chose 50 stem/derived pairs (900 in total) as test data. The remaining data were used as training items to estimate the parameters of the composition methods. Table 1 summarizes various characteristics of the dataset[3] (the last two columns of the table are explained in the next paragraphs).

Annotation of quality of test vectors The quality of the corpus-based vectors representing derived test items was determined by collecting human semantic similarity judgments in a crowdsourcing survey. In particular, we use the similarity of a vector to its nearest neighbors (**NNs**) as a proxy measure of quality. The underlying assumption is that a vector, in order to be a good representation of the meaning of the corresponding word, should lie in a region of semantic space populated by intuitively similar meanings, e.g., we are more likely to have captured the meaning of *car* if the NN of its vector is the *automobile* vector rather than *potato*. Therefore, to measure the quality of a given vector, we can look at the average similarity score provided by humans when comparing this very vector with its own NNs.

All 900 derived vectors from the test set were matched with their three closest NNs in our semantic space (see Section 4.2), thus producing a set of $2,700$ word pairs. These pairs were administered to CrowdFlower users,[4] who were asked to judge the relatedness of the two meanings on a 7-point scale (higher for more related). In order to ensure that participants were committed to the task and exclude non-proficient English speakers, we used 60 control pairs as gold standard, consisting of either perfect synonyms or completely unrelated words. We obtained 30 judgments for each derived form (10 judgments for each of 3 neighbor comparisons), with mean participant agreement of 58%. These ratings were averaged item-wise, resulting in a Gaussian distribution with a mean of 3.79 and a standard deviation of 1.31. Finally, each test item was marked as high-quality (**HQ**) if its derived form received an average score of at least 3, as low-quality (**LQ**) otherwise. Table 1 reports the proportion of HQ test items for each affix, and Table 2 reports some examples of HQ and LQ items with the corresponding NNs. It is worth observing that the NNs of the LQ items, while not as relevant as the HQ ones, are hardly random.

Annotation of similarity between stem and derived forms Derived forms differ in terms of how far their meaning is with respect to that of their stem. Certain morphological processes have systematically more impact than others on meaning: For example, the adjectival prefix *in-* negates the meaning of the stem, whereas *-ly* has the sole function to convert an adjective into an adverb. But the very same affix can affect different stems in different ways. For example, *remelt* means little more than to *melt again*, but *rethink* has subtler implications of changing one's way to look at a problem, and while one of the senses of *cycling* is present in *recycle*, it takes some effort to see their relation.

[3] Available from `http://clic.cimec.unitn.it/composes`

[4] `http://www.crowdflower.com`

Affix	Type	Derived form	Neighbors
-ist	HQ	transcendentalist	mythologist, futurist, theosophist
	LQ	florist	Harrod, wholesaler, stockist
-ity	HQ	publicity	publicise, press, publicize
	LQ	sparsity	dissimilarity, contiguity, perceptibility
-ment	HQ	advertisement	advert, promotional, advertising
	LQ	inducement	litigant, contractually, voluntarily
in-	HQ	inaccurate	misleading, incorrect, erroneous
	LQ	inoperable	metastasis, colorectal, biopsy
re-	HQ	recapture	retake, besiege, capture
	LQ	rename	defunct, officially, merge

Table 2: Examples of HQ and LQ derived vectors with their NNs

We conducted a separate crowdsourcing study where participants were asked to rate the 900 test stem/derived pairs for the strength of their semantic relationship on a 7-point scale. We followed a procedure similar to the one described for quality measurement; 7 judgments were collected for each pair. Participants' agreement was at 60%. The last column of Table 1 reports the average stem/derived relatedness (**SDR**) for the various affixes. Note that the affixes with systematically lower SDR are those carrying a negative meaning (*in-, un-, -less*), whereas those with highest SDR do little more than changing the POS of the stem (*-ion, -ly, -ness*). Among specific pairs with very low relatedness we encounter *hand/handy*, *bear/bearable* and *active/activist*, whereas *compulsory/compulsorily*, *shameless/shamelessness* and *chaos/chaotic* have high SDR. Since the distribution of the average ratings was negatively skewed (mean rating: 5.52, standard deviation: 1.26),[5] we took 5 as the rating threshold to classify items as having high (**HR**) or low (**LR**) relatedness to their stems.

4.2 Distributional semantic space[6]

We use as our source corpus the concatenation of ukWaC, the English Wikipedia (2009 dump) and the BNC,[7] for a total of about 2.8 billion tokens. We collect co-occurrence statistics for the top 20K content words (adjectives, adverbs, nouns, verbs) in lemma format, plus any item from the morphological dataset described above that was below this rank. The top 20K content words also constitute our context elements. We use a standard bag-of-words approach, counting collocates in a narrow 2-word before-and-after window. We apply (non-negative) Pointwise Mutual Information as weighting scheme and dimensionality reduction by Non-negative Matrix Factorization, setting the number of reduced-space dimensions to 350. These settings are chosen without tuning, and are based on previous experiments where they produced high-quality semantic spaces (Boleda et al., 2013; Bullinaria and Levy, 2007).

4.3 Implementation of composition methods

All composition methods except *mult* and *stem* have weights to be estimated (e.g., the λ parameter of *dilation* or the affix matrices of *lexfunc*). We adopt the estimation strategy proposed by Guevara (2010) and Baroni and Zamparelli (2010), namely we pick parameter values that optimize the mapping between stem and derived vectors directly extracted from the corpus. To learn, say, a *lexfunc* matrix representing the prefix *re-*, we extract vectors of *V/reV* pairs that occur with sufficient frequency (*visit/revisit*, *think/rethink*...). We then use least-squares methods to find weights for the *re-* matrix that minimize the distance between each *reV* vector generated by the model given the input *V* and the corresponding corpus-observed derived vector (e.g., we try to make the model-predicted *re+visit* vector as similar as possible to the corpus-extracted one). This is a general estimation approach that does not require task-specific hand-labeled data, and for which simple analytical solutions of the least-squares error prob-

[5]The negative skew is not surprising, as derived forms must have some relation to their stems!

[6]Most steps of the semantic space construction and composition pipelines were implemented using the DISSECT toolkit: `https://github.com/composes-toolkit/dissect`.

[7]`http://wacky.sslmit.unibo.it`, `http://en.wikipedia.org`, `http://www.natcorp.ox.ac.uk`

lem exist for all our composition methods. We use only the training items from Section 4.1 for estimation. Note that, unlike the test items, these have not been annotated for quality, so we are adopting an unsupervised (no manual labeling) but noisy estimation method.[8]

For the *lexfunc* model, we use the training items separately to obtain weight matrices representing each affix, whereas for the other models all training data are used together to globally derive single sets of affix and stem weights. For the *wadd* model, the learning process results in $0.16 \times \textit{affix} + 0.33 \times \textit{stem}$, i.e., the affix contributes only half of its mass to the composition of the derived form. For *dilation*, we stretch the stem (i.e., $\mathbf{v}$ of the dilation equation is the stem vector), since it should provide richer contents than the affix to the derived meaning. We found that, on average across the training pairs, *dilation* weighted the stem 20 times more heavily than the affix ($0.05 \times \textit{affix} + 1 \times \textit{stem}$). We then expect that the *dilation* model will have similar performance to the baseline *stem* model, as confirmed below.[9]

For all methods, vectors were normalized before composing both in training and in generation.

5 Experiment 1: approximating high-quality corpus-extracted vectors

The first experiment investigates to what extent composition models can approximate high-quality (HQ) corpus-extracted vectors representing derived forms. Note that since the test items were excluded from training, we are simulating a scenario in which composition models must generate representations for nonce derived forms.

Cosine similarity between model-generated and corpus-extracted vectors were computed for all models, including the *stem* baseline (i.e., cosine between stem and derived form). The first row of Table 3 reports mean similarities. The *stem* method sets the level of performance relatively high, confirming its soundness. Indeed, the parameter-free *mult* model performs below the baseline.[10] As expected, *dilation* performs similarly to the baseline, while *wadd* outperforms it, although the effect does not reach significance (p=.06).[11] Both *fulladd* and *lexfunc* perform significantly better than *stem* ($p < .001$). *Lexfunc* provides a flexible way to account for affixation, since it models it directly as a function mapping from and onto word vectors, without requiring a vector representation of bound affixes. The reason at the base of its good performance is thus quite straightforward. On the other hand, it is surprising that a simple representation of bound affixes (i.e., as vectors aggregating the contexts of words containing them) can work so well, at least when used in conjunction with the granular dimension-by-dimension weights assigned by the *fulladd* method. We hypothesize that these aggregated contexts, by providing information about the set of stems an affix combines with, capture the shared semantic features that the affix operates on.

	stem	*mult*	*dil.*	*wadd*	*fulladd*	*lexfunc*
All	0.47	0.39	0.48	0.50	0.56	0.54
HR	0.52	0.43	0.53	0.55	0.61	0.58
LR	0.32	0.28	0.33	0.38	0.41	0.42

Table 3: Mean similarity of composed vectors to high-quality corpus-extracted derived-form vectors, for all as well as high- (HR) and low-relatedness (LR) test items

When the meaning of the derived form is far from that of its stem, the *stem* baseline should no longer constitute a suitable surrogate of derived-form meaning. The LR cases (see Section 4.1 above) are thus crucial to understand how well composition methods capture not only stem meaning, but also affix-triggered semantics. The HR and LR rows of Table 3 present the results for the respective test subsets. As expected, the *stem* approach undergoes a strong drop when performance is measured on LR items. At the other extreme, *fulladd* and *lexfunc*, while also finding the LR cases more difficult, still clearly outperform the baseline ($p < .001$), confirming that they capture the meaning of derived forms beyond what their stems contribute to it. The effect of *wadd*, again, approaches significance when compared to the baseline ($p = .05$). Very encouragingly, both

[8] More accurately, we relied on semi-manual CELEX information to identify derived forms. A further step towards a fully knowledge-free system would be to pre-process the corpus with an unsupervised morphological induction system to extract stem/derived pairs.

[9] The other models have thousands of weights to be estimated, so we cannot summarize the outcome of parameter estimation here.

[10] This result does not necessarily contradict those of Mitchell and Lapata and others who found *mult* to be highly competitive. Due to differences in co-occurrence weighting schemes (we use a logarithmically scaled measure, they do not), their multiplicative model is closer to our additive one.

[11] Significance assessed by means of Tukey Honestly Significant Difference tests (Abdi and Williams, 2010)

	stem	*mult*	*wadd*	*dil.*	*fulladd*	*lexfunc*
-less	0.22	0.23	0.30	0.24	0.38	0.44
in-	0.39	0.34	0.45	0.40	0.47	0.45
un-	0.33	0.33	0.41	0.34	0.44	0.46

Table 4: Mean similarity of composed vectors to high-quality corpus-extracted derived-form vectors with negative affixes

fulladd and *lexfunc* significantly outperform *stem* also in the HR subset ($p<.001$). That is, the models provide better approximations of derived forms even when the stem itself should already be a good surrogate. The difference between the two models is not significant.

We noted in Section 4.1 that forms containing the "negative" affixes *-less*, *un-* and *in-* received on average low SDR scores, since negation impacts meaning more drastically than other operations. Table 4 reports the performance of the models on these affixes. Indeed, the *stem* baseline performs quite poorly, whereas *fulladd*, *lexfunc* and, to a lesser extent, *wadd* are quite effective in this condition as well, all performing greatly above the baseline. These results are intriguing in light of the fact that modeling negation is a challenging task for DSMs (Mohammad et al., 2013) as well as cDSMs (Preller and Sadrzadeh, 2011). To the extent that our best methods have captured the negating function of a prefix such as *in-*, they might be applied to tasks such as recognizing lexical opposites, or even simple forms of syntactic negation (modeling *inoperable* is just a short step away from modeling *not operable* compositionally).

6 Experiment 2: Comparing the quality of corpus-extracted and compositionally generated words

The first experiment simulated the scenario in which derived forms are not in our corpus, so that directly extracting their representation from it is not an option. The second experiment tests if compositionally-derived representations can be better than those extracted directly from the corpus when the latter is a possible strategy (i.e., the derived forms are attested in the source corpus). To this purpose, we focused on those 277 test items that were judged as low-quality (LQ, see Section 4.1), which are presumably more challenging to generate, and where the compositional route could be most useful.

We evaluated the derived forms generated by the models that performed best in the first experiment (*fulladd*, *lexfunc* and *wadd*), as well as the *stem* baseline, by means of another crowdsourcing study. We followed the same procedure used to assess the quality of corpus-extracted vectors, that is, we asked judges to rate the relatedness of the target forms to their NNs (we obtained on average 29 responses per form).

	corpus	*stem*	*wadd*	*fulladd*	*lexfunc*
All	2.28	3.26	4.12	3.99	3.09
HR	2.29	3.56	4.48	4.31	3.31
LR	2.22	2.48	3.14	3.12	2.52

Table 5: Average quality ratings of derived vectors

Target	**Model**	**Neighbors**
florist	*wadd*	flora, fauna, ecosystem
	fulladd	flora, fauna, egologist
	lexfunc	ornithologist, naturalist, botanist
sparsity	*wadd*	sparse, sparsely, dense
	fulladd	sparse, sparseness, angularity
	lexfunc	fragility, angularity, smallness
inducement	*wadd*	induce, inhibit, inhibition
	fulladd	induce, inhibition, mediate
	lexfunc	impairment, cerebral, ocular
inoperable	*wadd*	operable, palliation, biopsy
	fulladd	operable, inoperative, ventilator
	lexfunc	inoperative, unavoidably, flaw
rename	*wadd*	name, later, namesake
	fulladd	name, namesake, later
	lexfunc	temporarily, reinstate, thereafter

Table 6: Examples of model-predicted neighbors for words with LQ corpus-extracted vectors

The first line of Table 5 reports the average quality (on a 7-point scale) of the representations of the derived forms as produced by the models and baseline, as well as of the corpus-harvested ones (*corpus* column). All compositional models produce representations that are of significantly higher quality ($p < .001$) than the corpus-based ones. The effect is also evident in qualitative terms. Table 6 presents the NNs predicted by the three compositional methods for the same LQ test items whose corpus-based NNs are presented in Table 2. These results indicate that morpheme composition is an effective solution when the quality of corpus-extracted derived forms is low (and the previous experiment showed that, when their quality is high, composition can at least approximate corpus-based vectors).

With respect to Experiment 1, we obtain a different ranking of the models, with *lexfunc* being outperformed by both *wadd* and *fulladd* ($p<.001$), that are statistically indistinguishable. The *wadd*

composition is dominated by the stem, and by looking at the examples in Table 6 we notice that both this model and *fulladd* tend to feature the stem as NN (100% of the cases for *wadd*, 73% for *fulladd* in the complete test set). The question thus arises as to whether the good performance of these composition techniques is simply due to the fact that they produce derived forms that are near their stems, with no added semantic value from the affix (a "stemploitation" strategy).

However, the stemploitation hypothesis is dispelled by the observation that both models significantly outperform the *stem* baseline ($p<.001$), despite the fact that the latter, again, has good performance, significantly outperforming the corpus-derived vectors ($p < .001$). Thus, we confirm that compositional models provide higher quality vectors that are capturing the meaning of derived forms beyond the information provided by the stem.

Indeed, if we focus on the third row of Table 5, reporting performance on low stem-derived relatedness (LR) items (annotated as described in Section 4.1), *fulladd* and *wadd* still significantly outperform the corpus representations ($p<.001$), whereas the quality of the *stem* representations of LR items is not significantly different form that of the corpus-derived ones. Interestingly, *lexfunc* displays the smallest drop in performance when restricting evaluation to LR items; however, since it does not significantly outperform the LQ corpus representations, this is arguably due to a floor effect.

7 Conclusion and future work

We investigated to what extent cDSMs can generate effective meaning representations of complex words through morpheme composition. Several state-of-the-art composition models were adapted and evaluated on this novel task. Our results suggest that morpheme composition can indeed provide high-quality vectors for complex forms, improving both on vectors directly extracted from the corpus and on a stem-backoff strategy. This result is of practical importance for distributional semantics, as it paves the way to address one of the main causes of data sparseness, and it confirms the usefulness of the compositional approach in a new domain. Overall, *fulladd* emerged as the best performing model, with both *lexfunc* and the simple *wadd* approach constituting strong rivals. The effectiveness of the best models extended also to the challenging cases where the meaning of derived forms is far from that of the stem, including negative affixes.

The *fulladd* method requires a vector representation for bound morphemes. A first direction for future work will thus be to investigate which aspects of the meaning of bound morphemes are captured by our current simple-minded approach to populating their vectors, and to explore alternative ways to construct them, seeing if they further improve *fulladd* performance.

A natural extension of our research is to address morpheme composition and morphological induction jointly, trying to model the intuition that good candidate morphemes should have coherent semantic representations. Relatedly, in the current setting we generate complex forms from their parts. We want to investigate the inverse route, namely "de-composing" complex words to derive representations of their stems, especially for cases where the complex words are more frequent (e.g. *comfort/comfortable*).

We would also like to apply composition to inflectional morphology (that currently lies outside the scope of distributional semantics), to capture the nuances of meaning that, for example, distinguish singular and plural nouns (consider, e.g., the difference between the mass singular *tea* and the plural *teas*, which coerces the noun into a count interpretation (Katz and Zamparelli, 2012)).

Finally, in our current setup we focus on a single composition step, e.g., we derive the meaning of *inoperable* by composing the morphemes *in-* and *operable*. But *operable* is in turn composed of *operate* and *-able*. In the future, we will explore recursive morpheme composition, especially since we would like to apply these methods to more complex morphological systems (e.g., agglutinative languages) where multiple morphemes are the norm.

8 Acknowledgments

We thank Georgiana Dinu and Nghia The Pham for helping out with DISSECT-ion and the reviewers for helpful feedback. This research was supported by the ERC 2011 Starting Independent Research Grant n. 283554 (COMPOSES).

References

Hervé Abdi and Lynne Williams. 2010. Newman-Keuls and Tukey test. In Neil Salkind, Bruce Frey, and Dondald Dougherty, editors, *Encyclopedia of Research Design*, pages 897–904. Sage, Thousand Oaks, CA.

Harald Baayen, Richard Piepenbrock, and Leon Gulikers. 1995. The CELEX lexical database (release 2). CD-ROM, Linguistic Data Consortium, Philadelphia, PA.

Harald Baayen. 2005. Morphological productivity. In Rajmund Piotrowski Reinhard Köhler, Gabriel Altmann, editor, *Quantitative Linguistics: An International Handbook*, pages 243–256. Mouton de Gruyter, Berlin, Germany.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of EMNLP*, pages 1183–1193, Boston, MA.

Marco Baroni. 2009. Distributions in text. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics: An International Handbook*, volume 2, pages 803–821. Mouton de Gruyter, Berlin, Germany.

Laurie Bauer. 2001. *Morphological Productivity*. Cambridge University Press, Cambridge, UK.

Kenneth Beesley and Lauri Karttunen. 2000. *Finite-State Morphology: Xerox Tools and Techniques*. Cambridge University Press, Cambridge, UK.

Alan Black, Stephen Pulman, Graeme Ritchie, and Graham Russell. 1991. *Computational Morphology*. MIT Press, Cambrdige, MA.

Gemma Boleda, Marco Baroni, Louise McNally, and Nghia Pham. 2013. Intensionality was only alleged: On adjective-noun composition in distributional semantics. In *Proceedings of IWCS*, pages 35–46, Potsdam, Germany.

John Bullinaria and Joseph Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39:510–526.

Stephen Clark. 2012. Vector space models of lexical meaning. In Shalom Lappin and Chris Fox, editors, *Handbook of Contemporary Semantics, 2nd edition*. Blackwell, Malden, MA. In press.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis*, 36:345–384.

David Dowty. 1979. *Word Meaning and Montague Grammar*. Springer, New York.

Markus Dreyer and Jason Eisner. 2011. Discovering morphological paradigms from plain text using a Dirichlet process mixture model. In *Proceedings of EMNLP*, pages 616–627, Edinburgh, UK.

Katrin Erk. 2012. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.

Yoav Goldberg and Reut Tsarfaty. 2008. A single generative model for joint morphological segmentation and syntactic parsing. In *Proceedings of ACL*, pages 371–379, Columbus, OH.

John Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 2(27):153–198.

Sharon Goldwater and David McClosky. 2005. Improving statistical MT through morphological analysis. In *Proceedings of EMNLP*, pages 676–683, Vancouver, Canada.

Sharon Goldwater. 2006. *Nonparametric Bayesian Models of Lexical Acquisition*. Ph.D. thesis, Brown University.

Emiliano Guevara. 2009. Compositionality in distributional semantics: Derivational affixes. In *Proceedings of the Words in Action Workshop*, Pisa, Italy.

Emiliano Guevara. 2010. A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of GEMS*, pages 33–37, Uppsala, Sweden.

Graham Katz and Roberto Zamparelli. 2012. Quantifying count/mass elasticity. In *Proceedings of WCCFL*, pages 371–379, Tucson, AR.

Victor Kuperman. 2009. Semantic transparency revisited. Presentation at the 6th International Morphological Processing Conference.

Thomas Landauer and Susan Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

Young-Suk Lee. 2004. Morphological analysis for statistical machine translation. In *Proceedings of HLT-NAACL*, pages 57–60, Boston, MA.

Rochelle Lieber. 2004. *Morphology and Lexical Semantics*. Cambridge University Press, Cambridge, UK.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL*, pages 236–244, Columbus, OH.

Jeff Mitchell and Mirella Lapata. 2009. Language models based on semantic composition. In *Proceedings of EMNLP*, pages 430–439, Singapore.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

Saif Mohammad, Bonnie Dorr, Graeme Hirst, and Peter Turney. 2013. Computing lexical contrast. *Computational Linguistics*. In press.

Jason Naradowsky and Sharon Goldwater. 2009. Improving morphology induction by learning spelling rules. In *Proceedings of IJCAI*, pages 11–17, Pasadena, CA.

Ingo Plag. 1999. *Morphological Productivity: Structural Constraints in English Derivation*. Mouton de Gruyter, Berlin, Germany.

Anne Preller and Mehrnoosh Sadrzadeh. 2011. Bell states and negative sentences in the distributed model of meaning. *Electr. Notes Theor. Comput. Sci.*, 270(2):141–153.

Patrick Schone and Daniel Jurafsky. 2000. Knowledge-free induction of morphology using latent semantic analysis. In *Proceedings of the ConLL workshop on learning language in logic*, pages 67–72, Lisbon, Portugal.

Richard Socher, Eric Huang, Jeffrey Pennin, Andrew Ng, and Christopher Manning. 2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Proceedings of NIPS*, pages 801–809, Granada, Spain.

Richard Socher, Brody Huval, Christopher Manning, and Andrew Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of EMNLP*, pages 1201–1211, Jeju Island, Korea.

Richard Sproat. 1992. *Morphology and Computation*. MIT Press, Cambrdige, MA.

Peter Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Hsueh-Cheng Wang, Yi-Min Tien, Li-Chuan Hsu, and Marc Pomplun. 2012. Estimating semantic transparency of constituents of English compounds and two-character Chinese words using Latent Semantic Analysis. In *Proceedings of CogSci*, pages 2499–2504, Sapporo, Japan.

Richard Wicentowski. 2004. Multilingual noise-robust supervised morphological analysis using the wordframe model. In *Proceedings of SIGPHON*, pages 70–77, Barcelona, Spain.

David Yarowsky and Richard Wicentowski. 2000. Minimally supervised morphological analysis by multimodal alignment. In *Proceedings of ACL*, pages 207–216, Hong Kong.

Fabio Zanzotto, Ioannis Korkontzelos, Francesca Falucchi, and Suresh Manandhar. 2010. Estimating linear models for compositional distributional semantics. In *Proceedings of COLING*, pages 1263–1271, Beijing, China.