Massively Translingual Compound Analysis and Translation Discovery

Winston Wu, David Yarowsky
Department of Computer Science, Center for Language and Speech Processing
Johns Hopkins University
{wswu,yarowsky}@jhu.edu

Abstract

Word formation via compounding is a very widely observed but quite diverse phenomenon across the world’s languages, but the compositional semantics of a compound are often productively correlated between even distant languages. Using only freely available bilingual dictionaries and no annotated training data, we derive novel models for analyzing compound words and effectively generate novel foreign-language translations of English concepts using these models. In addition, we release a massively multilingual dataset of compound words along with their decompositions, covering over 21,000 instances in 329 languages, a previously unprecedented scale which should both productively support machine translation (especially in low resource languages) and also facilitate researchers in their further analysis and modeling of compounds and compound processes across the world’s languages.

Keywords: compounds, multilingual, translation

1. Introduction

Morphological compounding (e.g. lighthouse or airport) is one of the most common and productive methods of word formation across the world’s languages (Denning et al., 2007), and yet its derivational processes and semantics can be quite complex. Consider the semantic concept hospital, which can be realized via compound morphology in a remarkable diversity of semantic compositions, including:

| Lang. | Compound | Literal Semantics |
|-------|----------|-------------------|
| nl    | ziekenhuis | sick + house |
| no    | sykehus   | sick + house |
| hu    | kórház    | disease + house |
| eo    | malsanuelejo | sick + place |
| ms    | rumahsakit | house + sick |
| zh    | 病院      | disease + institution |

There are clearly a wide variety of semantic associations constituting this concept (e.g. sick/disease + house/place/institution), a variety of constituent orders (e.g. sick+house vs. house+sick) and potentially a variety of compounding processes beyond simple concatenation (e.g. sykehus in Norse (no) is a compound of syk and hus with the insertion of an e). The following paper presents a massively cross-linguistic computational model of both compound morphology compositional processes and compound semantics on a scale of over 300 languages.

Furthermore, the paper not only presents a derived analysis of the compounding process and semantics of compounds within a single language (e.g. German), as with much prior related work (e.g. Koehn and Knight (2003)), but does so via a joint model across essentially all the world’s languages with adequate dictionary resources, an unprecedentedly large scale for this class of research, and with significant additional synergistic multilingual power. In addition, the paper successfully applies these models and results to the valuable application of predicting novel translations of compound words, both to English (e.g. kórház → disease + house → hospital) and from English (e.g. hospital → disease + house, sick + place, etc. → kórház etc.), with valuable applications for translation dictionary expansion and out-of-vocabulary handling in machine translation, again on this uniquely large multilingual scale.

Finally, in conjunction with this paper, we release a novel and uniquely large-scale 329-language, 21,000+ example dataset1 of these compound morphological analyses and their associated compositional and compound translations, a valuable resource for training models for derivational morphology processes and compound semantics, with direct application to machine translation, on this massively multilingual scale.

2. Compound Discovery

We begin only with freely available multilingual translation dictionaries extracted from open-source Wiktionary2, with the hope that they contain both substantial examples of compounding in each language (e.g. sykehus (Norwegian) = hospital (English)) as well as translations of the constituents of these compounds (e.g. syk = sick and hus = house). Using these dictionaries, we develop a multi-iteration method for discovering both compound translation models or “recipes” motivated across multiple languages that can be used to construct or analyze new compound words that may not be in the dictionary. While there are many existing methods for compound splitting (e.g. Koehn and Knight (2003; Macherey et al. (2011)), we concern ourselves with compounds with two components, which can be combined by simple concatenation, optionally using a glue or filler string at the point of concatenation (Garera and Yarowsky, 2008), or by dropping the last character of the left component (henceforth drop-left). These three compounding processes productively cover a wide spectrum of our multilingual data and serve as an efficient foundation for training the semantic models of compounding in the absence of simple concatenation.

1github.com/wswu/worcomal
2www.wiktionary.org
In the first iteration of our method, we begin by considering only simple concatenation of two components that both exist in the dictionary (e.g. kör+ház = sick+house). By collecting words in all languages that can be decomposed into such components, we construct compound recipes as in Fig. 1. This process involves accounting for the varied order of components using a reordering and clustering method, augmenting the initial list of compounds in a second iteration of compound discovery by allowing glue characters and the drop-left mechanism, and scoring each word decomposition to indicate its validity as a compound word.

Throughout this paper, we will use the concept “hospital” as a running example. This is an interesting illustrative example since it is not a compound word in English, but occurs as a compound in many other languages.

### 2.1. Simple Concatenation

Many compound words can be discovered by simply splitting a string into all possible two parts and performing a dictionary lookup on each part. In fact, the large majority of compound words in our dataset are simple concatenations of legal stand-alone words. Table 1 presents a sample of such simple compound words. Note that concatenation can result in false positives (e.g. Dutch hospitaal = hospita ‘landlady’ + al ‘even’), which will be identified by the compound score described later.

### 2.2. Component Clustering

Within a compound word, the semantic ordering of the components often varies between languages. For example, compound words for the concept “hospital” have different component ordering in different languages:

- Dutch: zieken ‘sick’ + huis ‘house’
- Malay: rumah ‘house’ + sakit ‘sick’

To account for this variation, we cluster the components using a notion of syntagmatic and paradigmatic analysis. For each concept, we first filter out components that only occur once to remove noise. To illustrate, the top 5 left and right components for the concept “hospital”, before correcting for ordering, have the same components on both the left and right sides (Table 2). The numbers indicate the number of languages where we see that component on the left or right side, respectively.

### 2.3. Compound Validity Score

Not all words that can be decomposed into valid components are valid compound words. For example, the Dutch hospitaal (decomposed as hospita ‘landlady’ + al ‘even’) is clearly not a semantically meaningful compound word. To filter out these poor decompositions, we devise a measure of how well a compound’s components follow the compound’s recipe. A straightforward but effective score is the geometric mean of the highest counts of the left and right components, respectively. In situations where the a component is not in the recipe, it receives a language count of 0.1 (we do not use 0 because it will zero out the other term in the geometric mean). For example, the Hungarian kórházi (kór + ház; disease + house) receives a score of $\sqrt{8 \cdot 12} = 9.8$, indicating a good decomposition, while the Dutch “hospitaal” (hospita + al; landlady + even) receives a score of $\sqrt{0.1 \cdot 0.1} = 0.1$, indicating a bad decomposition. Scores for the concept “hospital” are presented in Table 3. Roughly
To construct new compound words, each pair of (foreign) concepts is concatenated using two new mechanisms: a glue character and drop-left mechanisms. We apply component clustering to score the new compound words as previously described. Note that recipes for one concept may result in a compound word for a new concept (e.g. in the second iteration, the Chinese concept ‘difficulty’ was constructed using the hospital recipe ‘ill’ (难受) + ‘place’ (处) using the drop-left mechanism).

Table 3: High-scoring decompositions for compounds of the concept "hospital".

| Score | Lng     | Decomposition              | Score | Lng     | Decomposition              |
|-------|---------|----------------------------|-------|---------|----------------------------|
| 12.96 | nn      | sjukhus = sjuk + hus; sick + house | 14    | house   | hospital +               |
| 12.96 | nl      | ziekenhuis = zieken + huis; sick + house | 11    | ill     | home                      |
| 12.96 | nl      | ziekenhuis = zieke + huis; sick + house | 8     | disease | building                 |
| 12.96 | no      | sykhus = syk + hus; sick + house | 7     | illness | place                     |
| 12.96 | no      | sykhus = syk + huis; sick + house | 5     | diseased | box                      |
| 12.96 | ms      | rumah sakit = rumah + sakit; house + sick | 4     | patient | family                    |
| 12.96 | sv      | sjukhus = sjuk + hus; sick + house | 3     | pain    | area                      |
| 12.96 | da      | sygehus = syge + hus; sick + house | 2     | ache    | case                      |
| 12.96 | da      | sygehus = syy + hus; sick + house | 2     | cure    | cottage                   |
| 12.96 | id      | rumah sakit = rumah + sakit; house + sick | 2     |         |                           |
| 12.96 | sv      | sjukhus = sjuk + hus; sick + house | 2     |         |                           |
| 12.96 | sv      | sjukhus = sjuke + hus; sick + home | 2     |         |                           |
| 10.58 | vi      | bệnh viện = bệnh + vien; sick + home | 2     |         |                           |
| 9.79  | hu      | kórház = kör + ház; disease + house | 14    | hospital | hospital +               |
| 9.79  | hu      | kórházi = kör + házi; disease + house | 11    | ill     | home                      |
| 6.93  | gd      | taigh-eiridinn = taigh + eiridinn; house + patient | 8     | disease | building                 |
| 6.48  | eo      | malsanulejo = malsanulo + ejo; sick + place | 7     | illness | place                     |
| 4.89  | eo      | taigh-leighis = taigh + leighis; house + heal | 5     | diseased | box                      |
| 4.00  | zh      | 病院 = 病院 + 院; disease + institution | 4     | patient | family                    |
| 4.00  | zh      | 病院 = 病院 + 院; disease + institution | 3     | pain    | area                      |
| 2.83  | zh      | 病院 = 病院 + 院; disease + institution | 2     | ache    | case                      |
| 2.83  | zh      | 病院 = 病院 + 院; disease + institution | 2     | cure    | cottage                   |
| 1.18  | tr      | hastane = hasta + ne; sick + en | 2     |         |                           |
| 1.18  | tg      | болничн = болничн + норн; sick + -land | 2     |         |                           |

Figure 2: Compounding recipe for the concept ‘hospital’ including glue character and drop left mechanisms.

We would like to see how well our methods work on compound words it has not seen before. Specifically, we evaluate our compounding methods on two tasks:

1. f2e: Given a foreign compound word, can we decompose and translate it into English?
2. e2f: Given an English concept, can we predict what the compound word would be in a target language?

For these experiments, we randomly chose a test set of 100 languages, with one word from each language that is likely to be a compound word according to our model. Due to the large multilingual breadth of this test set, evaluating on this test set gives a good idea of how the model performs on any given language of the world, rather than focusing on a single language with much more limited cross-linguistic generality. The randomly chosen test set is shown in Table 4. We remove each test word from the training dictionary to simulate it as being out-of-vocabulary.

4. Results and Analysis

For the e2f task, we were able to successfully recover 87 of the 100 test words. In other words, after removing the test word from the dictionary, the model was able to reconstruct the translation of the compound word only from its components, because other languages used the same components in the compounding recipe, either in direct association or via previously unobserved derived associations via the Kartesian product and/or reordering models. This result underscores that even if one does not observe a certain combination of components in any of the training data (e.g. sickness+building), the model’s inference that this semantic compounding is viable via model components of both reordering and semantic clustering of observed decompositions of other attested forms translating as hospital, facilitates our prediction that this novel association is more likely to occur and mean hospital if observed in monolingual target language data.
For the f2e task, we measured both accuracy and mean rank of the model’s translation hypotheses. Our method generated the correct English translation in 86/100 cases, a quite respectable performance given the great multilingual diversity of the test set, and the presence of frequently low resource languages where out-of-vocabulary compound words missing from the dictionary are quite common. Out of these cases, the correct English translation had a mean rank of 1.7 in the model’s ordered list of hypotheses, indicating that most of the time the correct English trans-
4.1. Dataset Analysis

Utilizing only simple direct concatenation, we were able to discover over 21,000 instances of English concepts that had were direct compounds of simpler constituents of known translation and explainable by one of the model’s recipes. By extending the modeled compounding mechanism to a single-character glue or dropping the last character of the left component, our method discovered an additional 2,700 concepts successfully analyzed as compounds.

4.2. Language-Specific Compounding Mechanisms

By examining the different processes used in constructing compound words, we obtain a greater understanding of how specific languages perform compounding. Table 5 shows stereotypical language-specific patterns. For example, most languages construct compounds simply by concatenating two words directly without insertions or deletions (although often in variable order). English often uses ‘i’ and ‘s’ as glue characters, while German uses ‘n’ and ‘s’. This information is not only useful for predicting whether a word is a compound, but can also be useful when generating previously unknown compound word translations into the language. A complete table and statistical analysis of these observed insertions and deletions in each language is included with our dataset, along with language-specific probabilities for the use of each type of compounding mechanism, a useful foundation for any compound-generating language model.

Certain languages like Chinese and Japanese are slightly problematic when discovering compounds using a character-dropping mechanism. For these languages, such a mechanism is not necessarily productive given that the dropped character is not merely a sound-insertion or basic-semantic-linking character but an important semantic component (that would not normally be associated with a single character in an alphabetic or even syllabic writing system). For instance, the Chinese 殺人 murder = 杀害 murder + 人 person is reasonable, but 音樂 music = 音律 tuning + 乐 music is not. Despite this caveat, single character dropping is still a widely observed and productive compounding process in these languages.

5. Related Work

Researchers have explored word compounding, though largely in the monolingual setting or on the order of a couple of languages. One multilingual effort similar to ours is MorBoComp (Guevara et al., 2006), a database of word compounds in 20 languages. The project seems to have stalled, and we were unable to access the data mentioned in their work. Our work encompasses a much larger set of languages (by a factor of 15x) and a much larger set of derived instances (even if their described database was actually available), and posits compound generation and analysis models absent from their work.

While we used straightforward but effective compound splitting algorithms, many more complicated splitting methods have been proposed, e.g. using n-gram counts (Sornlertlamvanich and Tanaka, 1996), supervised methods (Clouet and Daille, 2014), and monolingual and bilingual corpora (Koehn and Knight, 2003; Macherey et al., 2011) and could potentially employ in extensions of our work.

In contrast to several of these other works, the approach and analysis in our paper is simple yet effective in that it only
Table 5: Percentage of compound words in our dataset that were formed using the each compounding mechanism, along with common glue characters, if applicable.

| Lang | Concat | DL | Glue | Common Glues |
|------|--------|----|------|--------------|
| af   | 0.75   | 0.17| 0.08 | g, s         |
| br   | 0.85   | 0.12| 0.03 | o, r         |
| ca   | 0.79   | 0.16| 0.05 | a, l         |
| co   | 1.0    | 0   | 0    |              |
| com  | 0.5    | 0.5 | 0    |              |
| cop  | 0.27   | 0.73| 0    |              |
| crh  | 0.81   | 0.17| 0.01 | t, b         |
| cs   | 0.75   | 0.21| 0.04 | o, d         |
| dbh  | 1.0    | 0   | 0    |              |
| de   | 0.8    | 0.14| 0.06 | n, s         |
| dv   | 1.0    | 0   | 0    |              |
| ee   | 0.87   | 0.11| 0.02 | a            |
| el   | 0.65   | 0.35| 0    |              |
| eo   | 0.93   | 0.06| 0.01 | i, s         |
| es   | 1.0    | 0   | 0    |              |
| et   | 0.74   | 0.14| 0.12 | i, a         |
| eu   | 0.69   | 0.24| 0.06 | k, l         |
| fa   | 0.55   | 0.45| 0    |              |
| fax  | 1.0    | 0   | 0    |              |
| ff   | 0.86   | 0.61| 0.04 | n, s         |
| fy   | 0.5    | 0.42| 0.08 | l, t         |
| ha   | 0.5    | 0   | 0.5  | n, t         |
| haw  | 0.57   | 0.37| 0.06 | k, h         |
| ht   | 0.55   | 0.27| 0.18 | n, s         |
| hu   | 0.82   | 0.14| 0.04 | i, t         |
| ia   | 0.83   | 0.1  | 0.08 | i, l         |
| inh  | 0.42   | 0.24| 0.05 | a, n         |
| is   | 0.7    | 0.22| 0.08 | a, s         |
| ist  | 0      | 1.0 | 0    |              |
| it   | 0.77   | 0.18| 0.05 | s, r         |
| ja   | 0.32   | 0.68| 0    |              |
| jbo  | 0.21   | 0.05| 0.07 | n, r         |
| jv   | 0.82   | 0   | 0.18 | n, p         |
| ku   | 0.7    | 0.24| 0.05 | c, d         |
| kum  | 0      | 1.0 | 0    |              |
| kw   | 0.9    | 0.08| 0.02 | l             |
| ky   | 0.93   | 0.07| 0    |              |
| la   | 0.72   | 0.22| 0.06 | d, c         |
| lad  | 0.53   | 0.29| 0.18 | g, i         |
| lb   | 0.76   | 0.18| 0.06 | e, s         |
| lv   | 0.77   | 0.21| 0.02 | s, i         |
| pl   | 0.72   | 0.22| 0.05 | a, d         |
| prg  | 0      | 1.0 | 0    |              |
| pro  | 0.55   | 0.45| 0    |              |
| ps   | 0.57   | 0.43| 0    |              |
| pt   | 0.77   | 0.19| 0.04 | s, g         |
| qu   | 0.85   | 0.14| 0.01 | y, m         |
| raj  | 0      | 1.0 | 0    |              |
| rap  | 0.88   | 0   | 0.12 | r, m         |
| rm   | 0.45   | 0.43| 0.12 | r, g         |
| ro   | 0.77   | 0.19| 0.04 | r, i         |
| rup  | 0.5    | 0.45| 0.05 | c, t         |
| scn  | 0.55   | 0.33| 0.12 | n, g         |
| sco  | 0.48   | 0.43| 0.1  | n, g         |
| shn  | 0.17   | 0   | 0    |              |
| tpi  | 0.83   | 0.14| 0.03 | k, b         |
| tr   | 0.85   | 0.11| 0.04 | l, s         |
| uz   | 0.88   | 0.09| 0.03 | l, f         |
| vai  | 1.0    | 0   | 0    |              |
| vec  | 0.57   | 0.32| 0.11 | n, r         |
| vep  | 0.71   | 0.29| 0    |              |
| vi   | 0.74   | 0.25| 0.0  | t, n         |
| zh   | 0.34   | 0.66| 0    |              |

requires the usually very readily available on-line dictionaries in multiple languages (e.g. via Wiktionary or Panlex) without any analyzed seed training data. Because of this, our approach does not require potentially expensive linguistic annotation, and easily extends to multiple languages, as demonstrated compellingly by our successful scaling to 329 extremely diverse languages incorporating many morphological processes and character sets.

Translation of compound words using dictionaries have been explored by Garera and Yarowsky (2008). Our approach is similar in that we use multiple bilingual dictionaries, but we study and model the compounding phenomenon in more depth as well as on a much, much larger scale, with the significant benefits of much greater novel semantic pair discovery (both via direct observation and via our transitive cluster and reordering models). In addition, we release a very large 329-language 21,000+ instance large public resource of analyzed compound words and components and statistical analyses of their processes across all languages. In terms of applications, handling compound words well has been shown to improve machine translation, e.g. into English (Koehn and Knight, 2003) and German (Stymne et al., 2013) and has helped simplify medical text (Abrahamsson et al., 2014). We expect that our very large scale publicly distributed compound-based translation dictionaries and associated generative and analytic models will be useful for out-of-vocabulary handling in downstream machine translation systems, especially for low-resource languages.

6. Conclusion

While most languages exhibit broad-scale word formation via compounding, they often differ substantially in terms of the diverse processes by which words compound and novel concepts are realized via these compound processes. Using only freely available bilingual dictionaries and no annotated training data, we derived novel models for analyzing compound words and effectively generated novel foreign-language translations of English concepts using these models. In addition, we release a massively multilingual dataset of compound words along with their decompositions, covering over 21,000 instances in 329 languages, a previously unprecedented scale which we believe will both productively support machine translation (especially in low resource languages) and also facilitate researchers in their further analysis and modeling of compounds and compound processes across the world’s languages.

7. Acknowledgments

This work was supported in part by the DARPA LORELEI program. The findings, conclusions, and opinions found in this work are those of the authors and do not necessarily represent the views of the funding agency.

8. Bibliographical References

Abrahamsson, E., Forni, T., Skeppstedt, M., and Kvist, M. (2014). Medical text simplification using synonym replacement: Adapting assessment of word difficulty to a compounding language.
Clouet, E. and Daille, B. (2014). Splitting of compound terms in non-prototypical compounding languages.
Denning, K., Kessler, B., and Leben, W. R. (2007). *English vocabulary elements*. Oxford University Press.

Garera, N. and Yarowsky, D. (2008). Translating compounds by learning component gloss translation models via multiple languages. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Guevara, E., Scalise, S., Bisetto, A., and Melloni, C. (2006). Morbo/comp: a multilingual database of compound words. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06)*. European Language Resources Association (ELRA).

Koehn, P. and Knight, K. (2003). Empirical methods for compound splitting. In *10th Conference of the European Chapter of the Association for Computational Linguistics*.

Macherey, K., Dai, A. M., Talbot, D., Popat, A. C., and Och, F. (2011). Language-independent compound splitting with morphological operations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Volume 1*, pages 1395–1404. Association for Computational Linguistics.

Sornlertlamvanich, V. and Tanaka, H. (1996). The automatic extraction of open compounds from text corpora. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Stymne, S., Cancedda, N., and Ahrenberg, L. (2013). Generation of compound words in statistical machine translation into compounding languages. *Computational Linguistics*, 39(4).