Word Root Finder: a morphological segmentor based on CRF

Joseph Z. Chang *Jason S. Chang*
National Tsing Hua University, No. 101, Section 2, Kuang-Fu Road, Hsinchu, Taiwan
joseph.nthu.tw@gmail.com, jason.jschang@gmail.com

ABSTRACT
Morphological segmentation of words is a subproblem of many natural language tasks, including handling out-of-vocabulary (OOV) words in machine translation, more effective information retrieval, and computer assisted vocabulary learning. Previous work typically relies on extensive statistical and semantic analyses to induce legitimate stems and affixes. We introduce a new learning based method and a prototype implementation of a knowledge light system for learning to segment a given word into word parts, including prefixes, suffixes, stems, and even roots. The method is based on the Conditional Random Fields (CRF) model. Evaluation results show that our method with a small set of seed training data and readily available resources can produce fine-grained morphological segmentation results that rival previous work and systems.

KEYWORDS: morphology, affix, word root, CRF.

Proceedings of COLING 2012: Demonstration Papers, pages 51–58,
COLING 2012, Mumbai, December 2012.

1 Introduction

Morphological segmentation is the process of converting the surface form of a given word to the lexical form with additional grammatical information such as part of speech, gender, and number. The lexical form (or lemma) is the entries found in a dictionary or a lexicon. The conversion may involve stripping some prefixes or suffixes off the surface form.

For example, in *The Celex Morphological Database* (Baayen et al., 1996), the word *abstraction* is segmented into a stem *abstract* and a suffix *ion*. Celex provides additional grammatical information (e.g., the suffix *ion* in *abstraction*turns verb into noun. Our goal is to produce even more fine-grained segmentation, e.g., splitting the word *abstraction* into three meaningful units: *abs*, *tract* and *ion*, respectively meaning "*away*", "*draw*", and "*noun of verbal action*".

Constructing a fine-grained morphological system can potentially be beneficial to second language learners. Nation (2001) points out that an important aspect of learning vocabulary in another langauge is knowing how to relate unknown words and meanings to known word parts. English affixes and word roots are considered helpful for learning English. Understanding the meaning of affixes and roots in new words can expedite learning, a point emphasized in many prep books for standardized test such as GRE and TOEFL.

Many existing methods for morphological analysis rely on human crafted data, and therefore have to be redone for special domains. An unsupervised or lightly supervised method has the advantage of saving significant time and effort, when the need to adopt to new domains arises.

The problem can be approached in many ways. Most work in the literature focuses on inducing the morphology of a natural language, discovering the stems and affixes explicitly. An alternative approach is to build a morphological segmenter of words without having to produce a complete list of word parts including prefixes, suffixes, and stems (or roots).

The rest of the paper is organized as follows. In the next section, we survey the related work, and point out the differences of the proposed method. In Section 3, we describe in detail our method and a prototype system. Finally in Section 4, we report the evaluation results.

2 Related Work

Much research has investigated morphological analysis along the line of two level model proposed by Koskenniemi (1983). Recently, researchers have begun to propose methods for automatic analysis based on morphology knowledge induced from distributional statistics based on a corpus (Gaussier, 1999; Goldsmith, 1997) . In particular, Goldsmith (2001) shows that it is possible to generate legitimate stems and suffixes with an accuracy rate of 83% for English. More recently, Schone and Jurafsky (2000) propose to use word semantics from derived Latent Semantic Analysis (LSA) in an attempt to correct errors in morphology induction.

Morphological models or morphological segmenters can be used to keep the entries in a dictionary to a minimal by taking advantage of morphological regularity in natural language. Woods (2000) proposes a method that aggressively applies morphology to broaden the coverage a lexicon to make possible more conceptual and effective indexing for information retrieval. The author used around 1,200 morphological rules. Similarly, Gdaniec and Manandise (2002) show that by exploiting affixes, they can extend the lexicon of a machine translation system to cope with OOV words. We use a similar method to expand our seed training data.

More recently, Creutz and Lagus (2006) present Morfessor, an unsupervised method for segmenting words into frequent substrings that are similar to morphemes. The method is based on

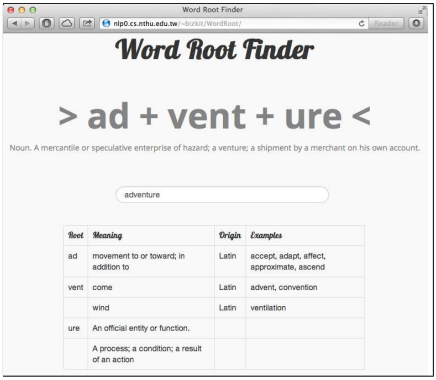

Figure 1: A system screen shot.

the principle of minimal description length (MDL), not unlike previous work such as Brent et al. (1995) and Goldsmith (2001). Additionally, Morfessor is enhanced by HMM states of *prefix*, *stems*, *suffix*, and *noise* based on morpheme length and successor/predecessor perplexity.

The system described in this paper differs from previous work in a number of aspects:

1. Previous work has focused mostly on two way splitting into stem and suffix (or amalgam of suffixes), while we attempt to split into Latin/Greek roots often found in English words.
2. We use a small set of words with hand annotation of prefixes, suffixes, and roots.
3. We experimented with several lists of affixes and a comprehensive lexicon (i.e., the Princeton WordNet 3.0) to expand the seed training data for better results.
4. We employ CRF with features from external knowledge sources to generalize from a small training set, without producing an explicit representation of morphology.

3 Method

In this section, we describe our method that comprises of three main steps. First, we automatically generate a training dataset by expanding a small set of seed annotated words (Section 3.1). In Step 2, we describe how to train a CRF model for word part segmentation (Section 3.2). Finally, we use the trained CRF model to construct a web-based system (Section 3.3).

3.1 Generate training data from seed data

To achieve reasonable coverage, supervised methods need a large training corpus. However, corpus annotated with fine-grained word parts is hard to come by. Here we describe two strategies that use a small set of annotated words to automatically generate a larger training set. The method is not unlike Woods (2000) or Gdaniec and Manandise (2002).

3.1.1 Expanding training data using prefix and suffix lists

Many words in English consist of stem, roots, and affixes. For examples, *finite* and *in+finite*, *senior* and *senior+ity*, *nation* and *inter+nation+al+ism*. Affix lists are not as difficult to come by,

comparing to word lists with fine-grained morphological annotations. With a list of affixes, we can iteratively and recursively attach prefixes and suffixes to words in the seed data, potentially forming a new annotated word. Since these expansions from a known word (e.g., *danger*) can be real words (e.g., *danger-ous*) as well as non-words (e.g., *danger-al*), we need to check each expansion against a dictionary to ensure the correctness. For example, with the list of affixes, "*in-, de-, -ness*", we can expand *fin+ite* into *fin+ite+ness*, *de+fin+ite*, *in+fin+ite*, *in+de+fin+ite*, *in+de+fin+ite+ness*.

3.1.2 Expanding training data using The Celex Database

Word lists annotated with more coarse-grained morphological annotations are also readily available, such as *The Celex Morphological Database*. Morphological annotations used in *The Celex Morphological Database* comprises of affixes and words, e.g., *abstract+ion*, while our target is to segment words into affixes and *word roots*, e.g., *abs+tract+ion*. By further segmenting the words in *The Celex Morphological Database* using the seed data, we can effectively generate more words for training. For example, with the seed word *abs+tract* and the *Celex* entry *abstract+ion*, we can successfully produce *abs+tract+ion*, an annotated word not found in the seed data.

3.2 Training a CRF model

After generating the training data, we treat each characters as a token, and generate several features using readily available affix lists. Our feature each token includes:

1. the character itself
2. whether the character is a vowel
3. does the remaining characters match a known suffix
4. does the preceding characters match a known prefix

We use two symbols for outcomes to represent segmentation: "+" indicates the character is the first character of the next word part, and "-" indicates otherwise. For example, if we want to segment the word *abstraction* into three parts: *abs*, *tract* and *ion*, the outcome sequence would be "- - - + - - - - + - -". Base on the generated features and annotations, we train a CRF model.

3.3 Runtime system

As the user of this system types in a word, the system continuously update the segmentation results on screen. A screen shot of our prototype [1] is shown in Figure 1, indicating that the user has entered the word *adventure*, and the system displays segmentation results, "*ad + vent + ure*", along with Wiktionary[2] definition. Additionally, information (based on Wiktionary and Wikipedia [3] of word parts, including definitions, origins, and examples are also displayed.

4 Evaluation and Discussion

We collected a total of 579 words (*Bennett-579*) with segmentation annotation from the book *Word Building with English Word Parts* by *Andrew E. Bennett* published by *Jong Wen Books Co.* in 2007. From the book, we randomly select 10%, or 60, annotated words for evaluation, identified in this paper as *Bennett-60*. The the remaining 90% forms a separate set of 519

[1] `morphology.herokuapp.com`
[2] `en.wiktionary.org`
[3] `en.wikipedia.org/wiki/List_of_Greek_and_Latin_roots_in_English` (as of Aug 22th, 2012)

training set	Bennett-60 test set			Bennett+XC-117 test set		
	tag prec.	tag rec.	word acc.	tag prec.	tag rec.	word acc.
Bennett-519	.85	.82	.80	.84	.57	.49
+XB	.88	.93	**.87**	.89	.87	.76
+XW	.81	.87	.78	.88	.80	.65
+XC	.85	.95	.83	.87	.80	.66
+XB+XW	.85	.87	.82	.89	.87	.76
+XB+XW+XC	.83	.87	.78	.92	.90	**.81**

Table 1: Evaluation results.

annotated words used for training, identified as *Bennett-519*. To more effectively evaluate the proposed method, we use *The Celex Morphology Database* with the method described in Section 3.1.2 to expand *Bennett-60* to *Bennett+XC-117* with 57 additional annotated words as the second test set. The Princeton WordNet 3.0 (Fellbaum, 1998) is used in the expansion process as a dictionary to ensure that the expanded words are legitimate.

Table 1 shows the evaluation results. We evaluate our system using three metrics: **tagging precision** and **tagging recall** indicate the tagging performance of the "+" tag. For example, if there are a total of 100 "+" tags in all outcome sequences, and the system tagged 50 tokens with the "+" tags, and 40 of them are correct. The tagging precision would be *80%*, and the tagging recall would be *40%*. **Word accuracy** is defined by the number of correctly tagged sequences, or words, divided by total number of test words. A sequence of outcomes for a word is considered correct, only when all the "+" and "-" tags are identical with the answer.

We explore the performance differences of using different resources to generate training data, the 6 systems evaluated are trained using the following traning sets repectively:

- **Bennett-519** : The system trained with the 519 annotated words from a book.
- **+XB** : A list of 3,308 annotated words expanded from **Bennett-519** with a list of 200 affixes collected from the same book.
- **+XW** : A list of 4,341 annotated words expanded from **Bennett-519** with a list of 1,421 affixes collected from Wikipedia.
- **+XC** : A list of 970 annotated words expanded by matching **Bennett-519** and *Celex*.
- **+XB+XW** : A list of 5,141 annotated words by combining **+XB** and **+XW**.
- **+XB+XW+XC** : A list of 5,366 annotated words by combining **+XB**, **+XW** and **+XC**.

As shown in Table 1, all six systems yield better performance on the *Bennett-60* test set than on the *Bennett+XC-117* test set, indicating the latter is a more difficult task. Further examining the two test sets, we found the everage number of segments per word is 2.7 for the *Bennett+XC-117* test set, and 2.0 for the *Bennett-60* test set. This is to be expected, since we generated *Bennett+XC-117* by extending words in *Bennet-60*. The **+XB** system performed the best on *Bennett-60*, with 87% word accurarcy. The **+XC** system ranked second, with 83% word accurarcy. For the *Bennett+XC-117* test set, the **+XB+XW+XC** system with all available traning data performed best with 81% word accurarcy, a 32% improvement comparing to the **Bennett-519** system trained using only the seed data.

In Tables 2 and 3, we list all 60 annotated words in the *Bennet-60* test set. The two tables respectively show the errornous/correct results of running **+XB** on the test set of *Bennett-60*.

By using supervised learning, we had to pay the price of preparing hand annotated training

answer	matri+x	sen+il+ity	ultra+violet	corp+se
result	matrix	senil+ity	ultra+vio+let	cor+pse
answer	loqu+acious	domi+cile	verit+able	mand+atory
result	loqu+aci+ous	dom+ic+ile	ver+it+able	mand+at+ory

Table 2: The 8 incorrect results and answers of running the **+XB** system on *Bennett-60* test set.

cycl+ist	endo+plasm	miss+ive	popul+ar	sub+scribe	with+stand
counter+point	dys+topia	milli+liter	poly+glot	son+ar	with+draw
con+fuse	doct+or	matri+mony	phonet+ics	sen+ior	voy+age
carn+al	dis+tort	lustr+ous	per+suade	se+cede	ver+ity
by+pass	dis+course	kilo+meter	patron+ize	re+tain	vent+ure
amphi+boly	dia+lect	in+pire	ob+struct	re+cline	tele+scope
ambi+ance	dextr+ity	hydr+ant	non+sense	pro+vide	tele+graph
de+flect	fin+ite	nomin+al	pre+view	sur+face	
de+cline	en+voy	nat+ion	pre+mature	super+vise	

Table 3: The 52 correct result of running the **+XB** system on *Bennett-60* test set.

data and lists of affixes, but we try to keep that to a minimum and used many existing resources to expand the dataset. However, the system does not require an internal lexicon at runtime and is capable of finding morphemes that is unseen in the training set and the affix lists. For example, many correcly identified morphemes shown in Table 3 such as *boly*, *topia*, *mony*, and *glot* are unseen morphemes. This shows by leveraging the set of rich features, the system provides a surprisingly high level of generality based on a relatively small training set.

Future work and summary

Many future research directions present themselves. We could handle cases where suffixes and words are not simply concatenated. For that, appending *ous* to *carnivore* should produces *carnivorous* instead of *carnivoreous*. A set of rules can be learned by using the manually annotated *Celex*. The same set of rules can also be used in runtime, to restore the segmented word roots to its original form. For example, after segmenting *advocation* into *ad+voc+at+ion*, we could modify *at+ion* into *ate+ion*, so that we can look up the meaning of the root *ate* in a affix dictionary. Additionally, an interesting direction to explore is incorporating more features in the CRF model. Statistics related to a prefix and the next letters (e.g., Prefix conditional entropy), or a suffix and preceding letter could be used as additional features in an attempt to improve accuracy. Yet another direction of research would be to disambiguate the meaning of affixes and roots, based on the definition or translation of the word, using known derivatives of affixes and word roots.

In summary, we have proposed a new method for constructing a fine-grained morphological word segmenter. The method comprises of three main parts, namely generating training data using a set of annotated seed data, generating features and label for training a CRF model for fine-grained word part segmentation, and a web-based prototype system. By combining two sets of manually annotated word lists, namely *Celex-2* and *Bennett-579*, we automatically produced enlarged training and test sets for more effective training and rigourous evaluation. Our system trained with all available training data is able to segment eight out of ten test words correctly. With the trained CRF model, we construct a web-base runtime system, a service that is potentially beneficial to English learners.

References

Baayen, R., Piepenbrock, R., and Gulikers, L. (1996). The CELEX Morphological Database, second edition.

Brent, M. R., Murthy, S. K., and Lundberg, A. (1995). Discovering morphemic suffixes a case study in MDL induction. In *The Fifth International Workshop on AI and Statistics*, pages 264–271.

Creutz, M. and Lagus, K. (2006). Morfessor in the morpho challenge. In *Proceedings of the PASCAL Challenge Workshop on Unsupervised Segmentation of Words into Morphemes*.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press.

Gaussier, E. (1999). Unsupervised learning of derivational morphology from inflectional lexicons. In *Proceedings of ACL Workshop on Unsupervised Learning in Natural Language Processing*.

Gdaniec, C. and Manandise, E. (2002). Using word formation rules to extend mt lexicons. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, AMTA '02, pages 64–73, London, UK. Springer-Verlag.

Goldsmith, J. (1997). *Unsupervised learning of the morphology of a natural language*. University of Chicago.

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

Koskenniemi, K. (1983). *Two-level morphology: a general computational model for word-form recognition and production*. Publications (Helsingin yliopisto. Yleisen kielitieteen laitos). University of Helsinki, Department of General Linguistics.

Nation, P. (2001). *Learning Vocabulary in Another Language*. The Cambridge Applied Linguistics Series. Cambridge University Press.

Schone, P. and Jurafsky, D. (2000). Knowledge-free induction of morphology using latent semantic analysis. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning - Volume 7*, ConLL '00, pages 67–72, Stroudsburg, PA, USA. Association for Computational Linguistics.

Woods, W. A. (2000). Aggressive morphology for robust lexical coverage. In *Proceedings of the sixth conference on Applied natural language processing*, ANLC '00, pages 218–223, Stroudsburg, PA, USA. Association for Computational Linguistics.