ACBiMA: Advanced Chinese Bi-Character Word Morphological Analyzer

Ting-Hao (Kenneth) Huang, Yun-Nung Chen, and Lingpeng Kong
Language Technologies Institute, Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213, USA
{tinghaoh, yvchen, lingpenk}@cs.cmu.edu

Abstract

While morphological information has been demonstrated to be useful for various Chinese NLP tasks, there is still a lack of complete theories, category schemes, and toolkits for Chinese morphology. This paper focuses on the morphological structures of Chinese bi-character words, where a corpus were collected based on a well-defined morphological type scheme covering both Chinese derived words and compound words. With the corpus, a morphological analyzer is developed to classify Chinese bi-character words into the defined categories, which outperforms strong baselines and achieves about 66% macro F-measure for compound words, and effectively covers derived words.

1 Introduction

Considering that Chinese is an analytic language without inflectional morphemes, Chinese morphology mainly focuses on analyzing morphological word formation. In this paper, we conceive the Chinese word forming process from a syntactic point of view (Packard, 2000). The analysis and prediction of the intra-word syntactic structures, i.e., the "morphological structures", have been shown to be effective in various Chinese NLP tasks, e.g., sentiment analysis (Ku et al., 2009; Huang, 2009), POS tagging (Qiu et al., 2008), word segmentation (Gao et al., 2005), and parsing (Li, 2011; Li and Zhou, 2012; Zhang et al., 2013). Thus, this paper focuses on analyzing the morphological structures of Chinese bi-character content words.

Huang et al. (2010) observed that 52% multi-character Chinese tokens are bi-character[1], which reflects that the core task of Chinese morphological analysis should be aimed at bi-character words. Previous work tended to focus on longer unknown words (Tseng and Chen, 2002; Tseng et al., 2005; Lu et al., 2008; Qiu et al., 2008) or the functionality of morphemic characters (Galmar and Chen, 2010), and none of them effectively covered Chinese bi-character words. To the best of our knowledge, Huang et al. (2010) is the only work focused on Chinese bi-character words, where they analyzed Chinese morphological types and developed a suite of classifiers to predict the types. However, their work covers only a subset of Chinese content words and has limited scalability. Therefore, this paper addresses the issues, which expands their work by developing a more detailed scheme and collecting more words to produce a generalized analyzer.

Our contributions are three-fold:

- Linguistic – we propose a morphological type scheme for full coverage of Chinese bi-character content words, and developed a corpus containing about 11K words.
- Technical – we develop an effective morphological classifier for Chinese bi-character words, achieving 66% macro F-measure for compound words, and and effectively covers derived words.
- Practical – we release the collected data and the analyzer with the trained model to provide additional Chinese morphological features for other NLP tasks. [2]

2 Morphological Type Scheme

Our morphological type category scheme is developed based on the literature (X.-H. Cheng, 1992; Lu et al., 2008; Huang et al., 2010) and the naming conventions of Stanford typed dependency (Chang

[1] The uni-character tokens do not contain any morphological structures.

[2] http://acbima.org/

Table 1: The category description and examples for derived words

Class	Morphological Characteristics	Example
dup	Two *duplicate* characters.	天天/tian-tian/day-day/everyday
pfx	The first character is a *prefix* character, e.g. 阿/a.	阿姨/a-yi/a-aunt/aunt
sfx	The second character is a *suffix* character, e.g. 仔/zi.	牛仔/new-zi/cow-zi/cowboy
neg	The first character is a *negation* character, e.g. 不/bu.	不能/bu-neng/no-capable/unable
ec	The first character is an *existential construction*, e.g. 有/you/have;exists.	有人/you-ren/exists-human/people

Table 2: The category description and examples for compound words

Class	Syntactic Role		Example
	Char 1	Char 2	
a-head	modifier	adjective head	最大/zui-da/most-big/biggest
n-head		nominal head	平台/ping-tai/flat-platform/(flat)platform
v-head		verbal head	主辦/zhu-ban/major-handle/host
nsubj	nominal subject	predicate (verb)	身經/shen-jing/body-experience/experience
vobj	predicate (verb)	object	開幕/kai-mu/open-screen/opening of event
vprt		particle	投入/tou-ru/throw-in to/throw in
conj	play coordinate roles in a word		男女/nan-nu/male-female/men and women (people)
els	else		transliterations, abbreviations, idiomatic words, etc.

et al., 2009; catherine De Marneffe and Manning, 2008) shown in Figure 1.

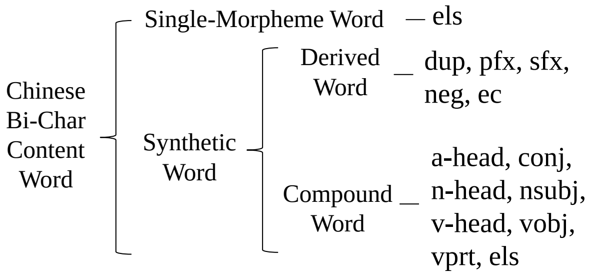

Figure 1: The morphological category scheme of Chinese bi-character content words

The two major categories of Chinese bi-character content words are *derived words* and *compound words*. Derived words are words formed in certain formations (e.g. duplication), while compound words are composed of constituent characters following certain syntactic relations. Table 1 and 2 present detailed category schemes. Note that for derived words, the characters "有/you/have" and "是/shi/be" are of a special type of existential constructions (Tao, 2007), so we isolate them from common prefixes to distinguish their unique characteristics. The "els" type (compound words) consists of exceptional words that cannot be categorized into our compound words scheme.

3 Morphological Type Classification

Due to the difference between derived words and compound words, we respectively adopt rule-based and machine learning approaches to predict their morphological types. Note that all of our approaches and features assume that Chinese morphological structures are independent from word-level contexts (Tseng and Chen, 2002; Li, 2011).

3.1 Derived Word: Rule-Based Approach

By definition, a morphological derived word can be recognized based on its formation. Therefore, we apply the pattern matching rules described in Table 1 to build a rule-based classifier.

To evaluate the coverage of these developed rules, we run the classifier on Chinese Treebank 7.0 (CTB) (Levy and Manning, 2003), where 2.9% of bi-character content words are annotated as derived words (842 unique word types). Our rules are able to capture derived words with a precision of 0.97. The false positives are caused by the ambiguity of Chinese characters "子/zi" and "兒/er".[3] The ambiguity results

[3] These two characters are common Chinese suffixes which mean "son/kid".

Table 3: Features for the Compound Word C1C2 (Dict: Revised Mandarin Chinese Dictionary (Ministry of Education (MoE), 1994); CTB: Chinese Treebank 5.1 (Xue et al., 2005))

Category		Feature	Description
Character Feature (for both C_i)	uni-char word	Tone	All possible tones (0-4) of C_i
		Pronunciation	All possible pronunciations, consonants, and vowels of C_i
		TF in CTB	The POS distribution of C_i in CTB
		Majority POS in CTB	The most frequent POS of C_i in CTB
		Character POS	Two POS tags when parsing the 2-token sentence C_1C_2
	uni-char morpheme	Dist. of Senses in Dict Majority POS in Dict	POS distribution of the senses of C_i in dictionary POS of C_i with the most senses in dictionary
	alphabet symbol	Root	The radical (also referred to as “character root”) of C_i
		CTB Prefix/Suffix Dist.	The occurrence distribution of the n-char words with C_i as the prefix/suffix corresponding to each POS in CTB.
		Dict Prefix/Suffix Dist.	The occurrence distribution of the n-char dictionary entry words with C_i as the prefix/suffix
		Example Word Prefix/Suffix Dist.	Same as above, but calculate the distribution in dictionary example words.
Word Feature (for C_1C_2)		Typed dependency	Typed dependency relation between C_1 and C_2
		Stanford Word POS	Single POS tag of a single token (word)

in mis-classifications such as “父子/fu-zi/father-son/father and son” into the “sfx” type instead of the “conj” type. Table 1 defines the patterns we consider as derived words, and the words that do not belong to the defined classes will be considered as compound words.

Table 4: Morphological category distribution

Category	Initial Set	Whole Set
	3,052 words	11,366 words
nsubj	1.2%	1.6%
v-head	7.7%	8.7%
a-head	1.1%	1.8%
n-head	36.7%	34.0%
vprt	9.4%	9.3%
vobj	14.3%	14.6%
conj	25.5%	26.9%
els	4.1%	3.3%

3.2 Compound Word: Machine Learning Approach

To automatically predict morphological types for compound words, we perform machine learning techniques to capture generalizations from various features. For each bi-character word C_1C_2, we extract *character-level* features for C_1 and C_2 individually, as well as a single *word-level* feature for C_1C_2. Table 3 describes our feature set. For character-level features, a Chinese character may take on 3 different roles: word, morpheme, or alphabet symbol, where the extracted features are organized according to these roles. In addition, we propose word-level features, e.g. POS of C_1C_2, to capture the word information dismissed by the previous work (Huang et al., 2010) with consideration that such clue helps classification.

We experiment with various ML classification models: Naïve Bayes (John and Langley, 1995), Random Forest (Breiman, 2001), and Support Vector Machine (Platt, 1999; Keerthi et al., 2001; Hastie and Tibshirani, 1998) for the classification task. The three types of baselines are compared: Majority, Stanford Dependency Map, and Tabular Models. The Tabular Models first assign the POS tags to each known character C based on different heuristics (i.e., the most frequent POS of C in CTB, the POS of C with most senses in Dict, and the POS of C annotated by Stanford Parser), and then assigns the most frequent morphological type obtained from training data to each POS combination, e.g., “(VV, NN) = vobj”. The Stanford Dependency Map takes the dependency relation between C_1 and C_2 as predicted by the Stanford Parser (Chang et al., 2009) , and maps it to a corresponding morphological type, which is learned from training data. The Majority baseline always outputs the majority type, i.e., the “n-head” type.

Table 5: 10-fold cross-validation classification performance (MF: Macro F-measure, ACC: Accuracy)

Approach	nsubj	v-head	a-head	n-head	vprt	vobj	conj	els	MF	ACC
Majority	0	0	0	.507	0	0	0	0	.172	.340
Stanford Dep. Map	0	0	0	.525	.351	.438	.213	.010	.332	.388
Tabular (Stanford POS)	0	.296	0	.524	.389	.434	.162	.064	.349	.395
Tabular (CTB POS)	.021	.337	.009	.645	.397	.529	.421	.095	.479	.508
Tabular (Dict POS)	0	.292	.060	.670	.253	.572	.494	.035	.495	.526
Naïve Base	.273	.406	.195	.523	.679	.566	.547	.188	.519	.518
Random Forest	.250	.421	.063	**.760**	**.803**	.643	**.656**	.076	.647	**.674**
SVM	**.413**	**.541**	**.288**	.748	.791	**.657**	.636	**.271**	**.662**	.665
Avg Difficulty Level	1.74	1.55	1.64	1.36	1.38	1.38	1.47	1.95	-	-

4 ACBiMA Corpus 1.0

We develop a Chinese morphological type corpus containing 11,366 bi-character compound words, referred to as "ACBiMA Corpus 1.0." This corpus is incrementally developed in two stages:

The "initial set" is first developed for preliminary study and analysis. We randomly extracted about 3,200 content words from Chinese Treebank 5.1 (Xue et al., 2005), and removed the derived words. After manually checking for and removing errors, the initial set contains 3,052 words, which are further annotated with "morphological types" and "difficulty level of determining" (1, 2, or 3) by trained native speakers and examined again by experts. The inter-annotator agreement on a 50-word held-out set, averaged over all annotator pairs, is 0.726 Kappa.

In the second stage, we expand on the initial set into a larger corpus for practical use. We sampled about 3,000 words from CTB 5.1 and annotated them with their morphological types. Moreover, we obtained the 6,500-word corpus developed by Huang et al. (2010)[4] and manually split its "Substantive-Modifier" words into "a-head", "n-head", or "v-head" types to match our category scheme. In total, the expanded dataset consists of 11,366 unique bi-character compound word types (see Table 4).

5 Experiments

We performed 10-fold cross-validation experiments on the entire dataset to evaluate our approaches for compound words.[5] As mentioned in §3.2, we compared against different baselines. Table 5 presents the results of our experiments, and the average human-judged difficulty level (in initial set) is also listed for comparison.

Random Forest and SVM outperformed all other models and baselines. The best accuracy is 0.674; 65% of words in the initial set are labeled as "easy" by human annotators, suggesting that our classifiers are comparable to human performance on the "easy" instances. Also, we achieved similar level of performance in macro F1-measure when compared to Huang et al. (2010)[6], despite our task being more challenging due to having two extra types.

6 Conclusion and Future Work

In this paper, we developed a set of tools and resources for leveraging morphology of Chinese bi-character words. We propose a category scheme, develop a corpus, and build an effective morphological analyzer. In future work, we intend to explore other NLP tasks where we can take advantage of ACBiMA and our tools to improve performance.

Acknowledgments

We thank anonymous reviewers for their useful comments. We are also grateful to Yanchuan Sim for his helpful feedback and all participants who helped to annotate the data.

[4] The words in Huang et al. (2010) are sampled from the NTCIR CIRB040 news corpus, and the distribution of types is similar to that of our initial set. This suggests that the morphological types distribution between different Chinese corpora are similar.

[5] For the 3 machine learning algorithms, we used the implementations found in the Weka toolkit (Hall et al., 2009).

[6] They reported macro F1-measure of 0.67.

References

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Marie catherine De Marneffe and Christopher D. Manning, 2008. *Stanford typed dependencies manual.*

Pi-Chuan Chang, Huihsin Tseng, Dan Jurafsky, and Christopher D. Manning. 2009. Discriminative reordering with chinese grammatical relations features. In *Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation*, SSST '09, pages 51–59, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bruno Galmar and Jenn-Yeu Chen. 2010. Identifying different meanings of a chinese morpheme through semantic pattern matching in augmented minimum spanning trees. *Prague Bull. Math. Linguistics*, 94:15–34.

Jianfeng Gao, Mu Li, Andi Wu, and Chang-Ning Huang. 2005. Chinese word segmentation and named entity recognition: A pragmatic approach. *Comput. Linguist.*, 31(4):531–574, December.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November.

Trevor Hastie and Robert Tibshirani. 1998. Classification by pairwise coupling. *The Annals of Statistics*, 26(2):451–471, 04.

Ting-Hao Huang, Lun-Wei Ku, and Hsin-Hsi Chen. 2010. Predicting morphological types of chinese bi-character words by machine learning approaches. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Ting-Hao Huang. 2009. Automatic extraction of intra- and inter- word syntactic structures for chinese opinion analysis. Master's thesis, Graduate Institute of Networking and Multimedia, National Taiwan University.

George H. John and Pat Langley. 1995. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, pages 338–345, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy. 2001. Improvements to platt's smo algorithm for svm classifier design. *Neural Comput.*, 13(3):637–649, March.

Lun-Wei Ku, Ting-Hao Huang, and Hsin-Hsi Chen. 2009. Using morphological and syntactic structures for chinese opinion analysis. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, pages 1260–1269, Stroudsburg, PA, USA. Association for Computational Linguistics.

Roger Levy and Christopher Manning. 2003. Is it harder to parse chinese, or the chinese treebank? In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 439–446, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhongguo Li and Guodong Zhou. 2012. Unified dependency parsing of chinese morphological and syntactic structures. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 1445–1454, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhongguo Li. 2011. Parsing the internal structure of words: A new paradigm for chinese word segmentation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1405–1414, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jia Lu, Masayuki Asahara, and Yuji Matsumoto. 2008. Analyzing chinese synthetic words with tree-based information and a survey on chinese morphologically derived words. In *IJCNLP'08*, pages 53–60.

Taiwan Ministry of Education (MoE). 1994. Revised mandarin chinese dictionary. Online Version. Available at http://dict.revised.moe.edu.tw.

Jerome L. Packard, 2000. *The Morphology of Chinese: A Linguistic and Cognitive Approach*, chapter 3.1.1.4. Cambridge University Press, New York.

John C. Platt. 1999. Advances in kernel methods. chapter Fast Training of Support Vector Machines Using Sequential Minimal Optimization, pages 185–208. MIT Press, Cambridge, MA, USA.

Likun Qiu, Changjian Hu, and Kai Zhao. 2008. A method for automatic pos guessing of chinese unknown words. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 705–712, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hongyin Tao. 2007. Subjectification and the development of special-verb existential/presentative constructions. *Language and Linguistics*, 8(2):575–602.

Huihsin Tseng and Keh-Jiann Chen. 2002. Design of chinese morphological analyzer. In *Proceedings of the First SIGHAN Workshop on Chinese Language*

Processing - Volume 18, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

Huihsin Tseng, Daniel Jurafsky, and Christopher Manning. 2005. Morphological features help pos tagging of unknown words across language varieties.

X.-L. Tian. X.-H. Cheng. 1992. *Modern Chinese*. Bookman Books Ltd.

Naiwen Xue, Fei Xia, Fu-dong Chiou, and Marta Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Nat. Lang. Eng.*, 11(2):207–238, June.

Meishan Zhang, Yue Zhang, Wanxiang Che, and Ting Liu. 2013. Chinese parsing exploiting characters. In *ACL (1)*, pages 125–134. The Association for Computer Linguistics.