ACBiMA: Advanced Chinese Bi-Character Word Morphological Analyzer

Ting-Hao (Kenneth) Huang
Yun-Nung (Vivian) Chen
Lingpeng Kong

HTTP://ACBIMA.ORG

Outline

- Introduction
- Related Work
- Morphological Type Scheme
- Morphological Type Classification
  - Drived Word: Rule-Based Approach
  - Compond Word: ML Approach
- Experiments
  - ACBiMA Corpus 1.0
  - Experimental Results
- Conclusion & Future Work

Outline

- **Introduction**
- Related Work
- Morphological Type Scheme
- Morphological Type Classification
  - Drived Word: Rule-Based Approach
  - Compond Word: ML Approach
- Experiments
  - ACBiMA Corpus 1.0
  - Experimental Results
- Conclusion & Future Work

Introduction

- NLP tasks usually focus on segmented words
- **Morphology** is how words are composed with morphemes
- Usages of Chinese morphological structures
  - Sentiment Analysis (Ku, 2009; Huang, 2009)
  - POS Tagging (Qiu, 2008)
  - Word Segmentation (Gao, 2005)
  - Parsing (Li, 2011; Li, 2012; Zhang, 2013)
- Challenge for Chinese morphology
  - Lack of complete theories
  - Lack of category schema
  - Lack of toolkits

−	−	+
抗	菌	
anti	bacteria	
verb	*object*	

Outline

- Introduction
- **Related Work**
- Morphological Type Scheme
- Morphological Type Classification
  - Drived Word: Rule-Based Approach
  - Compond Word: ML Approach
- Experiments
  - ACBiMA Corpus 1.0
  - Experimental Results
- Conclusion & Future Work

Related Work

- Focus on *longer unknown words*
  - Tseng, 2002; Tseng, 2005; Lu, 2008; Qiu, 2008
- Focus on the *functionality* of morphemic characters
  - Bruno, 2010
- Focus on *Chinese bi-character words*
  - Huang, et al., 2010 (LREC)
    - 52% multi-character Chinese tokens are bi-character
    - analyze Chinese morphological types
    - developed a suite of classifiers for type prediction

Issue: covers only a subset of Chinese content words and has limited scalability

Outline

- Introduction
- Related Work
- **Morphological Type Scheme**
- Morphological Type Classification
  - Drived Word: Rule-Based Approach
  - Compond Word: ML Approach
- Experiments
  - ACBiMA Corpus 1.0
  - Experimental Results
- Conclusion & Future Work

Morphological Type Scheme

- Chinese Bi-Char Content Word
  - Single-Morpheme Word — els
  - Synthetic Word
    - Derived Word — dup, pfx, sfx, neg, ec
    - Compound Word — a-head, conj, n-head, nsubj, v-head, vobj, vprt, els

Derived Word

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

Compond Word

Class	Syntactic Role		Example
	Char 1	Char 2	
a-head n-head v-head	modifier	adjective head nominal head verbal head	最大/zui-da/most-big/biggest 平台/ping-tai/flat-platform/(flat)platform 主辦/zhu-ban/major-handle/host
nsubj	nominal subject	predicate (verb)	身經/shen-jing/body-experience/experience
vobj vprt	predicate (verb)	object particle	開幕/kai-mu/open-screen/opening of event 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.

Outline

- Introduction
- Related Work
- Morphological Type Scheme
- **Morphological Type Classification**
  - **Drived Word: Rule-Based Approach**
  - **Compond Word: ML Approach**
- Experiments
  - ACBiMA Corpus 1.0
  - Experimental Results
- Conclusion & Future Work

Morphological Type Classification

- Assumption: Chinese morphological structures are independent from word-level contexts (Tseng, 2002; Li, 2011)
- Derived words
  - Rule-based approach
- Compound words
  - ML-based approach

Derived Word: Rule-Based

- Idea
  - a morphologically derived word can be recognized based on its formation
- Approach
  - pattern matching rules
- Evaluation
  - Data: Chinese Treebank 7.0
  - Result:
    - 2.9% of bi-char content words are annotated as derived words
    - Precision = 0.97

Rule-based methods are able to effectively recognizing derived words.

Compond Word: ML-Based

- Idea
  - The characteristics of individual characters can help decide the type of compond words
- ML classification models
  - Naïve Bayes
  - Random Forest
  - SVM

Classification Feature

- Dict: Revised Mandarin Chinese Dictionary (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	POS distribution of the senses of C_i in dictionary
		Majority POS in Dict	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)

Outline

- Introduction
- Related Work
- Morphological Type Scheme
- Morphological Type Classification
  - Drived Word: Rule-Based Approach
  - Compond Word: ML Approach
- **Experiments**
  - **ACBiMA Corpus 1.0**
  - **Experimental Results**
- Conclusion & Future Work

ACBiMA Corpus 1.0

- Initial Set
  - 3,052 words
  - Extracted from CTB5
  - Annotated with difficulty level
- Whole Set
  - 11,366 words
  - Initial Set +
    3k words from CTB 5.1 +
    6.5k words from (Huang, 2010)

Table 4: Morphological category distribution

Category	Initial Set 3,052 words	Whole Set 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%

Baseline Models

1) Majority
2) Stanford Dependency Map
3) Tabular Models
   - Step 1: assign the POS tags to each known character based on different heuristics
   - Step 2: assign the most frequent morphological type obtained from training data to each POS combination, e.g., “(VV, NN) = vobj”

Experimental Result

- Setting: 10-fold cross-validation
- Metrics: Macro F-measure (MF), Accuracy (ACC)

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	0	.296	0	.524	.389	.434	.162	.064	.349	.395
	CTB	**.021**	**.337**	.009	.645	**.397**	.529	.421	**.095**	.479	.508
	Dict	0	.292	**.060**	**.670**	.253	.572	**.484**	.035	**.495**	**.526**

Tablular approaches perform better among all baselines.

Experimental Result

- Setting: 10-fold cross-validation
- Metrics: Macro F-measure (MF), Accuracy (ACC)

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	0	.296	0	.524	.389	.434	.162	.064	.349	.395
	CTB	.021	.337	.009	.645	.397	.529	.421	.095	.479	.508
	Dict	0	.292	.060	.670	.253	.572	.484	.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

ML-based methods outperform all baselines, where SVM & RF perform best.

Experimental Result

- Setting: 10-fold cross-validation
- Metrics: Macro F-measure (MF), Accuracy (ACC)

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	0	.296	0	.524	.389	.434	.162	.064	.349	.395
	CTB	.021	.337	.009	.645	.397	.529	.421	.095	.479	.508
	Dict	0	.292	.060	.670	.253	.572	.484	.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		1.74	1.55	1.64	1.36	1.38	1.38	1.47	1.95	-	-

Outline

- Introduction
- Related Work
- Morphological Type Scheme
- Morphological Type Classification
  - Drived Word: Rule-Based Approach
  - Compond Word: ML Approach
- Experiments
  - ACBiMA Corpus 1.0
  - Experimental Results
- **Conclusion & Future Work**

Conclusion & Future Work

- Contribution
  - Linguistic
    - Propose a morphological type scheme
    - Develop a corpus containing about 11K words
  - Technical
    - Develop an effective morphological classifier
  - Practical
    - Data and tool available
    - Additional features for any Chinese task
- Future
  - Improve other NLP tasks by using ACBiMA

Q & A

Thanks for your attentions!!

- HTTP://ACBIMA.ORG