Chinese Lexical Resources for Gender, Number, Animacy

Zhiyi Song, Jiahong Yuan, Xiaoyi Ma, Stephanie Strassel

Linguistic Data Consortium. University of Pennsylvania
3600 Market Street, Suite 810, Philadelphia, PA. 19104 USA
{zhiyi, jiahong, xma, strassel} AT ldc.upenn.edu

1. Introduction

In this paper, we report on an effort to build Chinese lexicons on gender, number, and animacy for DARPA's Deep Exploration and Filtering of Text (DEFT) program, following the method described in Ji and Lin (2009). These resources are intended to support improved performance for DEFT performers of core NLP capabilities like entity tagging. .

Gender, number, and animacy are lexical indicators that can be useful in the detection of person mentions. Ji and Lin (2009) demonstrated that the probabilistic lexical properties of gender and animacy could be applied to detect person mentions from English raw texts in an unsupervised manner, with results comparable to state-of-the-art supervised learning methods. In their study, the gender and animacy properties of English words were learned from counting noun-gender and noun-animacy pairs in web-scale 5-grams. However, the queries employed in Ji and Lin (2009) for counting co-occurrences in English texts cannot and should not be simply adapted to Chinese, due to the nature of the syntactic structure of Chinese. For example, a noun phrase and its co-referential pronoun may often be separated by more than five words in Chinese. So rather than using 5-grams, we opted to use full sentences as te target range to discover co-occurances in Chinese. In addition, Chinese-specific constructions such as the noun phrase structure, "BA" (把字句: result of an action) and 'BEI" (被字句: passive construction) constructions can provide useful information for discovering the lexical property of animacy, and should be included in the queries.

We used the Chinese Gigaword corpus (5th edition, LDC2011T13), which consists of approximately 3.09 billion characters. The Chinese lexicons were derived from querying full sentences in the corpus, as described in Sections 2 and 3. The following procedures were used to develop the Chinese lexicons:

1. Sentence segmentation: we first segmented all source documents into sentences based on punctuation marks.
2. Traditional Chinese conversion: All source documents in traditional Chinese script were converted to simplified Chinese script.
3. POS tagging: A POS tagger (Xue, 2013) was run on all sentences.
4. Query development: We developed Chinese-specific queries to extract patterns based on POS tags and Chinese syntactic features (Bergsma, 2005; Bergsma and Lin, 2006; Ji and Lin, 2009).
5. Lexicon building: We then counted the number of times each common noun or proper noun phrase co-occurred with gender or number pronouns, and also counted entity types along the extracted patterns, to extract gender, number and animacy lexicons.

2. Queries for Gender and Number Lexicons

We developed four queries for building gender and number lexicons (the POS tags used in the examples are defined in Xia, 2000):

2.1. Conjunction-possessive

In this query, we intended to extract all cases like

(1) 阿拉法特_NR 和_CC 他_PN 的_DEG 随员_NN
Arafat and his companions

(2) 总统_NN 与_CC 他_PN 的_DEG 顾问_NN
The president and his consultant

in which the proper noun (NR) or common noun (NN) in the conjunction is highly likely to be co-referenced with the pronoun, so that we can rely on the number and gender of the pronoun to predict the number and gender of the proper noun or common noun phrase. This query yielded 9,037 proper noun entries and 4,960 common noun entries.

2.2. Nominative-predicate

In this query, we intended to extract all cases like

(3) 他_PN 是_VC 北韩_NR 历来_AD 访_VV 美_NR 的_DEC 最_JJ 高级_JJ 官员_NN

He is the highest ranked North Korean official who visited US

in which the noun phrase (NN) and the pronoun (PN) are coreferential (官员= 他) but exclude cases like

(4) 他_PN 是_VC 在_P 昨天_NT 遭_VV 子弹_NN 击中_VV 头部_NN
It was yesterday that he suffered a bullet attack in head

in which 是_VC is for focus, comparison or emphasis and the noun phrase and the pronoun do not corefer (头部 =\= 他). We also exclude cases like

(5) 他们_PN 是_VC 施哈柏_NR 和_CC 马福德_NR
They are Schewad and Marford

in which the pronoun is the coreference of conjunctions of noun phrases. By filtering out strings that contain {在}_P or{和}_CC from the extracted results, we were able to exclude unwanted patterns. We then relied on the number and gender of the pronouns to predict the number and gender of the proper noun or common noun phrases. This query yielded 79,872 proper noun entries and 20,623 common noun entries.

2.3. Verb-nominative

In this query, we intended to extract all cases like

(6) 贝克汉姆_NR 1 日_NT 表示_VV ,_PU 他_PN 最近_NT 两_CD 天_M 拜会_VV 了_AS
Bechham indicated on 1st that he visited …

in which the noun phrase (NR or NN) and the pronoun are most likely coreferential(贝克汉姆=他), but exclude cases like

(7) 国际足联_NR 应该_VV 终止_VV 他_PN 的_DEG 陈述_NN
FIFA should terminate his statement

in which the noun phrase and the pronoun most likely do not corefer (国际足联=\=他). By restricting the verbs to speech and cognition verbs, eg, 表示 indicate, 宣布 announce, 打算 plan etc..

We were able to further extract cases in which the pronouns are most likely to be coreferential with noun phrases in the subject position. We then relied on the number and gender of the pronouns to predict the number and gender of the proper noun or common noun phrase. This query yielded 79,872 proper noun entries and 20,623 common noun entries.
The Chinese pronouns used for the three queries above include:

Gender and number	**Code**	**Pronouns**
Feminine Plural	FP	她们 *they*
Feminine Singular	FS	她 *she, her*
Inanimate Plural	IP	它们 *they*
Inanimate Singular	IS	它 *it*
Masculine Singular	MS	他 *he, him*
Neuter Plural	NP	他们, 我们, 大家, 你们, 双方 *they, them, we, us, you, all, both*
Neuter Singular	NS	自己, 我, 你, 自身, 各自, 本身, 对方, 本人, 自己, 自我 *I, you, self, each, other*

Table 1: Pronouns for Gender and Number

2.4. Gender designator

In this query, we intended to extract two proper noun and common noun combinations:

- Apposition structure (NNNR) in which the first word is a common noun NN and the second is a proper noun NR, eg.
  (8) 英国_NR 王子_NN 威廉_NR *William, Prince of England*
- Proper noun and common noun combination (NRNN) in which the first word is a proper noun NR and second is a common noun NN), eg.
  (9) 李蓉_NR 女士_NN *Ms. Li Rong*

In both cases, the common nouns can be used to predict the gender of the proper nouns, similar to the way Mr. or Mrs. function in English. We manually identified and verified a list of gender designators (including masculine, feminine and neutral) from each structure. Next, we counted the frequencies of each proper noun in association with these gender designators and then derived a dictionary entry for each proper noun that included a likely gender categorization based on

its gender designator frequency distribution. The NNNR structures yielded 981,468 proper noun entries and the NRNN structure yielded 94,961 proper noun entries.

3. Queries for Animacy Lexicons

As described in Ji and Lin (2009), the English dictionary for animacy relies on relative pronouns as a cue. Since there is no overt relative pronoun in Chinese, we had to seek other approaches to obtain animacy lexicons.

3.1. BA Structure

Generally, Chinese has a tendency to prefer animate nouns in the subject position, but this is not always the case. Other Chinese comprehension research studies (Phillip et al, 2008; Zhang, 2001) indicate that BA and BEI constructions in Chinese have a strong tendency to have an animate subject, especially in formal writing. Following that line of thinking, we extracted subject noun phrases containing BA and BEI constructions, relying on those constructions to serve as predictors of animacy of Chinese nouns. There are not many instances of BEI construction in the Chinese Gigaword corpus, so we only derived lexicons from the BA construction. For example, in the clause

(10) 俄罗斯_NR 外交部_NN 当天_NT 将_BA 此_PN 决定_VV 通知_VV 了_AS …
Russian Foreign Ministry notified this decision to

(11) 两_CD 名_M 嫌疑人_NN 把_BA 一_CD 个_M 用药瓶_NN 制成_VV 的_DEC 小型_JJ 炸弹_NN 扔进_VV …
Two suspects threw a mini bomb made from drug container into …

外交部 (*Foreign Ministry*) in (10) and 嫌疑人(*suspects*) in (12) are the subjects of the BA constructions, and they were extracted as animate noun subjects. This query yielded 39,110 proper noun entries and 69,615 common noun entries.

3.2. Proper noun and common noun combinations

Appositions such as 总理_NN 温家宝_NR (*PM Wen Jiabao*) are abundant in the Chinese Gigaword corpus. By restricting NN to title designators, we can then safely assume that the proper nouns in the apposition are persons (mostly singular), which are animate. We manually identified titles from the most frequent common nouns in apposition structure and supplemented the list with Chinese title lists shared by other researchers. This query yielded 1,034,387 person names.

Additionally, the meaning of common nouns in the apposition structure (NNNR) and the combination of proper noun and common noun structure (NRNN) can be used to predict the entity type of the proper nouns in these two structures. For example, in the apposition structure,

(12) 总理_NN 温家宝_NR (*PM Wen Jiabao*)

总理 indicates that 温家宝 is a person entity;

(13) 执政党_NN 社会党_NR (*Socialist Party, the party in power*)

执政党 indicates that 社会党 is an organization entity. Similarly, in compound noun phrases (NR_NN),

(14) 斯里兰卡_NR 东北部_NN (*Northeast of Sri Lanka*)

东北部 can be used to predict that 斯里兰卡 is a location entity;

(15) 兰德_NR 研究所_NN (*Lander Institute*)

研究所 can predict that 兰德研究所 is a proper noun of an organization entity. We manually identified designators for person, organization and location entities from the most frequent common nouns in NR_NN structures and derived lexicons for each entity type. Altogether, the apposition structure yielded 2,312,202 named entities.

4. Conclusion

In summary, below are the queries that we used to derive Chinese gender, number and animacy dictionaries from Chinese Gigaword 5th edition.

Property	Proper noun	Target (entries)	Context	Example
Gender and Number	Conjunction-Possessive	Noun (4,960) Proper noun (9,037)	Pronoun in Conjunction	**阿拉法特**和**他**的… (Arafat and his…)
	Nominative-Predicate	Noun (15,826) Proper noun (3,198)	Pronoun with Copula	**他**是**老板** (he is a boss)
	Verb-Nominative	Noun (20,623) Proper noun (79,872)	Pronoun with Verb	**贝克汉姆** 1 日表示, **他**最近... (Beckham indicated that he…)

	Gender-designator	Proper noun (1,076,429)	Designator	克林□先生 (Mr. Clinton)
Animacy	Apposition	Proper noun (1,034,387)	Title	□理**温家宝** (Prime minster Wen Jiabao)
	BA construction	Noun (69,615) Proper noun (39,110)	Ba Construction	**嫌疑人**将炸□扔□学生当中 (suspects threw a bomb)
	Apposition	Named Entity (1,018,719)	Designator	成□国**委内瑞拉** (Venezuela, a member country)
	proper noun and common noun combo	Organization Named Entity (77,906)	Designator	**□德研究所** (Lander Institute)
	proper noun and common noun combo	Named Entity (196,858)	Designator	**席**博士 (Doctor Xi)

Table 2: Chinese dictionaries for Gender, Number and Animacy

5. Acknowledgements

This material is based on research sponsored by the Air Force Research Laboratory and Defense Advanced Research Projects Agency under agreement number FA8750-13-2-0045. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory and Defense Advanced Research Projects Agency or the U.S. Government.

The authors thank Heng Ji for sharing her Chinese title list, Xiaopeng Bai for sharing his Chinese noun and verb taxonomy annotation .

6. References

Shane Bergsma. Automatic Acquisition of Gender Information for Anaphora Resolution. Proceedings of Canadian Artificial Intelligence, 2005 and http://www.clsp.jhu.edu/~sbergsma/Pubs/Presentations/bergsmaCanAI2005.publish.pdf

Shane Bergsma and Dekang Lin. Bootstrapping Path-Based Pronoun Resolution, In Proceedings of the Conference on Computational Linguistics / Association for Computational Linguistics (COLING/ACL-06), Sydney, Australia, July 17-21, 2006

Heng Ji and Dekang Lin. Gender and Animacy Knowledge Discovery from Web-Scale N-Grams for Unsupervised Person Mention Detection. http://www.aclweb.org/anthology/Y/Y09/Y09-1024.pdf, 2009

Robert Parker, et al. Chinese Gigaword Fifth Edition LDC2011T13. Web Download. Philadelphia: Linguistic Data Consortium, 2011

Markus Philipp, Ina Bornkessel-Schlesewsky, Walter Bisang, Matthias Schlesewsky. The role of animacy in the real time comprehension of Mandarin Chinese: Evidence from auditory event-related brain potentials. Brain and Language 105 (2008) 112–133. 2008

Nianwen Xue. Chinese word segmentation as character tagging. Computational Linguistics and Chinese Language Processing 8.1: 29-48. 2003

Xia, F. (2000) The Part-of-Speech Guidelines for Chinese Treebank Project. Technical Report IRCS 00-07, University of Pennsylvania.

Baijiang Zhang. 被字句和把字句的对称与不对称. 中国语文. 2001 年第 6 期