Chinese Word Segmentation in FTRD Beijing

Heng LI
France Telecom R&D Beijing
heng.li@francetelecom.com

Yuan DONG
France Telecom R&D Beijing
yuan.dong@francetelecom.com

Xinnian MAO
France Telecom R&D Beijing
xinnian.mao@francetelecom.com

Haila WANG
France Telecom R&D Beijing
haila.wang@francetelecom.com

Wu LIU
Beijing University of Posts and Telecommunications
wu.liu@francetelecom.com.cn

Abstract

This paper presents a word segmentation system in France Telecom R&D Beijing, which uses a unified approach to word breaking and OOV identification. The output can be customized to meet different segmentation standards through the application of an ordered list of transformation. The system participated in all the tracks of the segmentation bakeoff -- PK-open, PK-closed, AS-open, AS-closed, HK-open, HK-closed, MSR-open and MSR-closed -- and achieved the state-of-the-art performance in MSR-open, MSR-close and PK-open tracks. Analysis of the results shows that each component of the system contributed to the scores.

1 Introduction

The development of the Chinese word segmentation system presented in this bakeoff began in Feb. this year, and will last for one year with the support of the ILAB Beijing initial project within France Telecom R&D.

Although the project last only half year by now, the main components of the system has been implemented, including code identification and conversion, basic segmentation, factoid detection, morphological analysis, name entity identification, segmentation standards adaptor, except the components of code identification and conversion and segmentation standards adaptors, other components are integrated in a statistical framework of n-gram language model.

2 System Description

2.1 Code identification and conversion

For processing both Simplified and Traditional Chinese text from a variety of locales, including Mainland China, Hong Kong and Taiwan, we choose UTF-8 as internal character representation within the system. The ability to transparently handle Chinese text from any Chinese locale greatly simplifies the logic of the segmentation system.

2.2 N-gram language model

In our system, Chinese words can be categorized into one of the following types: lexicon words, morphological words, factoids, name entities. These types of words are processed in different ways in our system, and are incorporated into a unified statistical framework of the trigram language model.

2.2.1 Basic segmentation

Each input sentence is first segmented into individual characters. These characters and the character strings are then looked up in a lexicon. For the efficient search, the lexicon is represented by a TRIE compressed in a double-array data struc-

ture. Given a character string, all its prefix strings that form lexicon words can be retrieved efficiently by browsing the TRIE whose root represents its first character.

2.2.2 Factoid detection

There are twenty four kinds of factoid words, such as time, date, money, etc. All the factoid words are represented as regular expressions, and compiled into a compressed DFA with the row-index algorithm.

2.2.3 Morphological analysis

As (Wu 2003) discussed in the paper, it is those morphologically derived words (MDWs hereafter) that are most controversial and most likely to be treated differently in different standards and different systems. In our system, there are six main categories of morphological processes, affixation, directional verb, resultative verb, splitting verb, reduplication and merging, and we employ a chart parsing algorithm augmented with word lattices structure which incorporates the morphological rules especially designed for Chinese languages with restrictive CFG.

2.2.4 Name entity identification

Our NE identification concentrates on three types of NEs, namely, personal names (PERs), location names (LOCs) and organization names (ORGs). For Chinese person names, we only consider PN candidates that begin with a family name stored in the family name list and follow a given name which is of one or two characters long. For transliterations of foreign person names, a PN candidate would be generated if it contains only characters stored in a transliterated character list. For location names and organizations names, we only use the LN list and ON list to generate the candidates.

2.3 Segmentation standards adaptor

In this bakeoff, there are four segmentation standards and slightly different from ours. Standard adaptation is conducted with the application of an ordered list of transformations on the output of our segmentation system. The method we use is Transformation-Based Learning, and the transformation templates are lexicalized templates. In our system, we designed 14 lexicalized templates.

2.4 Speed

As we optimized our lexicon and decoding process, the speed of segmentation is very fast. On a single 2.80 GHz, 1G bytes memory, Xeon machine, the system is able to process about 0.73 Mega bytes per second.

The speed may vary according to the sentence lengths: given texts of the same size, those containing longer sentences will take more time. The number reported here is an average of the time taken to process the test sets of the eight tracks we participated in.

3 Evaluation

3.1 Open tracks

In the open tracks, we used four lexicons of 210,319 entries, 165,103 entries, 174,268 entries, 165,655 entries respectively on AS-open, HK-open, MSR-open, PK-open tracks, which include the entries of 2,430 MDWs, 12,487 PNs, 22,907 LNs and 29,032 ONs, 10,414 four-character idioms, plus the word lists generated from the training data provided by the bakeoff. We use the training data provided by the bakeoff for training our trigram word-based language model. We also used a family name list (which contains 399 entries in our system), and a 1,021-entry transliterated name character list.

3.2 Closed tracks

In the close tracks, the lexicon we use could only be generated from the training data provided by the bakeoff. We could only use the training data provided by the bakeoff for training our word-based language model. Also, since the training data we used is only from the bakeoff, there does not exist any different standards, standards adaptor component is not necessarily needed.

3.3 Result analysis

Our system is designed so that components such as the factoid detection and NE identification can be switched on or off, so that we can investigate the relative contribution of each component to the overall word segmentation performance. The results are summarized in the table 1. For comparison, we also include in the table (Row 1) the results of using FMM. Row 2 shows the baseline results of our system, where only the lexicon is used. Each cell in the table has six fields. From the top, there are respectively Precision, Recall, F-measure, OOV Recall, IV Recall and Speed (Mega bytes/second). We don't list the speed in Row 6 since it decreases a factor of 10 to 60 because of application of thousands of TBL rules.

	PK_o	PK_c	MSR_o	MSR_c	AS_o	AS_c	HK_o	HK_c
1. FMM	0.857	0.841	0.921	0.917	0.871	0.864	0.842	0.838
	0.925	0.906	0.968	0.957	0.925	0.911	0.928	0.908
	0.891	0.872	0.945	0.936	0.898	0.887	0.885	0.872
	0.143	0.069	0.107	0.025	0.097	0.014	0.175	0.162
	0.947	0.957	0.971	0.982	0.947	0.952	0.961	0.968
	2.435	2.570	2.951	3.090	2.813	2.937	2.748	2.850
2. Baseline	0.869	0.855	0.931	0.926	0.891	0.877	0.863	0.851
	0.941	0.928	0.973	0.969	0.943	0.942	0.930	0.929
	0.905	0.890	0.952	0.947	0.917	0.908	0.897	0.888
	0.235	0.069	0.275	0.025	0.132	0.014	0.194	0.162
	0.960	0.987	0.987	0.995	0.982	0.984	0.985	0.990
	0.967	1.017	0.879	0.923	0.703	0.728	0.921	0.956
3. 2+FT	0.946	0.919	0.950	0.940	0.903	0.900	0.873	0.862
	0.951	0.950	0.973	0.973	0.945	0.947	0.932	0.932
	0.948	0.934	0.961	0.956	0.924	0.923	0.902	0.895
	0.748	0.448	0.396	0.205	0.180	0.156	0.292	0.215
	0.963	0.980	0.990	0.944	0.979	0.983	0.983	0.989
	0.819	0.879	0.779	0.787	0.631	0.635	0.821	0.830
4. 3+MA	0.946	0.919	0.950	0.940	0.903	0.900	0.873	0.862
	0.951	0.950	0.973	0.973	0.945	0.947	0.932	0.932
	0.948	0.934	0.961	0.956	0.924	0.923	0.902	0.895
	0.748	0.448	0.371	0.205	0.181	0.156	0.295	0.215
	0.963	0.980	0.989	0.944	0.979	0.983	0.983	0.989
	0.807	0.879	0.753	0.787	0.626	0.635	0.815	0.830
5. 4+NE	0.951	0.919	0.956	0.940	0.920	0.900	0.900	0.862
	0.957	0.950	0.973	0.973	0.949	0.947	0.938	0.932
	0.954	0.934	0.965	0.956	0.934	0.923	0.918	0.895
	0.788	0.448	0.454	0.205	0.330	0.156	0.411	0.215
	0.967	0.980	0.956	0.944	0.977	0.983	0.980	0.989
	0.679	0.879	0.716	0.787	0.604	0.635	0.748	0.830
6. 5+adaptation	0.960	0.919	0.957	0.940	0.919	0.900	0.901	0.862
	0.964	0.950	0.975	0.974	0.952	0.948	0.940	0.932
	0.962	0.934	0.966	0.957	0.935	0.923	0.920	0.895
	0.788	0.449	0.453	0.210	0.311	0.158	0.410	0.215
	0.974	0.980	0.989	0.995	0.981	0.983	0.982	0.989

Table 1. Our system results on all the tracks.

From Table 1 we can find that, in rows 1 and 2, the dictionary-based methods already achieve quite good recall, but the precisions are not very good because they cannot correctly identify unknown words that are not in the lexicon such as factoids and name entities. We also find that even using the same lexicon, our approach that is based on the N-gram language models outperforms the greedy approach because the use of context model resolves more ambiguities in segmentation. As shown in Rows 3 to 5, when components are switched on in turn, the overall word segmentation performance increases consistently. The morphological analysis has no contribution to the overall performance in Row 4. The main reason is that the number of MDWs used in our system is very small (only 2,430) and there may exist very small MDWs in the test sets. The similar cases occur on NE identification in the close tracks in Row 5 since we would not do NE identification at all in the close tracks. We also notice that the contribution of NE identification is very little in the open tracks, which shows that the performance of NE identification is not very good in our system, and explains why our OOV recall is not very high compared

with other participants in the bakeoff. This is one area of our future work to improve. The results of standards adaptation on four bakeoff test sets are shown in Row 6. It turns out that performance except IV recall improves slightly across the board in all four test sets. The main reason is that the training data and lexicon we used are mainly from the four providers in the bakeoff, there does not exist any different segmentation standards.

4 Conclusions

The evaluation results show that the closed tests is not very good compared with other participants, the one main reason is that the word-based language model we used is not competitive compared with other algorithms in the closed tracks. One area of our future work is to apply other machine learning algorithm, like Maximum Entropy (ME), Support Vector Machine (SVM), Conditional Random Field (CRF), etc.

Acknowledgements

The work reported here was a team effort. We thank Wu Liu, Haitao Zeng, Nan He for their help in the experimentation and evaluation of the system.

References

Andi Wu. 2003. Customizable segmentation of morphologically derived words in Chinese. *International Journal of Computational Linguistics and Chinese Language Processing,* 8(1): 1-27.

Aoe, J. 1989. An Efficient Digital Search Algorithm by Using a Double-Array Structure. *IEEE Transactions on Software Engineering,* Vol. 15, 9: 1066-1077.

George Anton Kiraz. 1999. Compressed Storage of Sparse Finite-State Transducers. *4th International Workshop on Automata Implementation*, Pages: 109-121.

Jian Sun, Ming Zhou and Jianfeng Gao. 2003. Chinese named entity identification using class-based language model. *International Journal of Computational Linguistics and Chinese Language Processing,* 8(1).

Jianfeng Gao, Mu Li, Andi Wu and Chang-Ning Huang. 2004a. Chinese word segmentation: a pragmatic approach. Microsoft Research Technical Report, MSR-TR-2004-123.

Julia Hockenmaier, Chris Brew. 1998. Error driven segmentation of Chinese. *Communications of COLIPS,* 8(1): 69-84.

Xinnian Mao, Heng Li, Yuan Dong, Haila Wang. 2005. Chinese Morphological Analyzer. *IEEE NLP-KE 2005*, submitted.