High Coverage Lexicon for Japanese Company Name Recognition

Xu Liang　Yasufumi Taniguchi　Hiroki Nakayama
TIS Inc.　TIS Inc.　TIS Inc.
{ryo.sho, taniguchi.yasufumi, nakayama.hiroki}@tis.co.jp

1 Introduction

Company name recognition is a curial task for constructing enterprise knowledge graph in the financial domain. A straightforward approach is utilizing a company lexicon for recognition. Such company lexicons are usually collected from web sources, which cause two problems, the long tail company name recognition problem and the variant forms recognition problem. The long tail company name recognition problem means the company names collected from web usually only contain a small portion of total company names. For example, there are 4 million registered companies in Japan, but most of company names in web source are related to listed companies, which only has 3,704 companies[1]. In other words, most of long tail company names are not covered. The second problem is the variant forms recognition problem, which means a company's name can have different forms, such as the legal form, the colloquial form and so on. The colloquial form may contain acronym, person names, locations, numbers, and other unusual tokens. This problem becomes more challenging in the Japanese scenario because a Japanese company's name can compose four kinds of different notation systems, kanji, hiragana, katakana, and latin characters.

In order to solve these problems, we collect company names from Japan National Tax Agency instead of from the web, and then propose an alias generation workflow to generate alias as much as possible. Finally, we cat get a high coverage lexicon (Japanese Company Lexicon: JCL[2]) to cover long tail company names and different variant forms of company names.

In the experiments, we compare JCL with other widely used Japanese lexicons (IPAdic, NEologd, Juman) from intrinsic perspective (coverage test) and extrinsic perspective (downstream NLP task) in two datasets. The experiment results shows that JCL has highest coverage and can be used for coarse grained annotation to avoid human annotation cost.

2 Japanese Company Lexicon Generation

In this section, we generate the Japanese Company Lexicon (JCL) with two steps, collecting company names from lexicon source, and generate aliases with rules.

2.1 Lexicon Source

In order to cover long tail companies as much as possible, we use the open data from the National Tax Agency, which is the official tax collecting agency of Japan. The National Tax Agency assigns a corporate number to each corporation and open the basic information. The basic information for each corporation contains the unique corporate number, the registered address, and the trade name.

We collect total 4,794,402 trade names. But we only consider the "Company" types that the organizations conducts economic activity for commercial purposes. The "Company" types are denoted as 株式会社 (Stock Compay), 有限会社 (Limited Company), and 合同会社 (Limitted Liability Company). The total number of the three company types are 4,140,409 (2,258,329, 1,684,516, and 197,564, respectively), such as TIS株式会社 (TIS Inc.), はごろもフーズ株式会社 (Hagoromo Foods Corporation), DOWAホールディングス株式会社 (DOWA HOLDINGS Co., Ltd), 株式会社ティ・シィ・エス (TCS CO., LTD.) and so on.

2.2 Alias Generation

For the alias generation process, we mainly pay attention to the patterns of Japanese company names. If a company name only contain kanji, the alias usually only has two form, for example, the 株式会社紀伊國屋書店 and 紀伊國屋書店. If a company name contains latin characters, it usually also has the katakana form. For example, the company TIS

[1] Untill 2020/01/09

[2] https://github.com/BrambleXu/Japanese-Company-Lexicon

株式会社 could be represented as ティーアイエス株式会社.

The alias generation process consists of the 6 steps, and we use 株式会社ザ・レジェンド for the demonstration (see Table 1).

	Step	Example
1	Add English name	The Legend Co.,Ltd.
2	Add katakana name	ザ・レジェンド
3	Remove legal form designations	ザ・レジェンド, The Legend
4	Remove special character	ザレジェンド
5	Normalization	ザレジエンド
6	Remove duplicates	[株式会社ザ・レジェンド, ザレジェンド, ザレジエンド, ザ・レジェンド, The Legend Co.,Ltd.,The Legend]

Table 1: Alias generation process

Add English name and katakana name Some Japanese companies have the English name or the katakana name in the open data. We add these names if they are available.

Remove legal form designations Company names are usually represented as colloquial names in web resources. For example, we usually use TIS instead of TIS 株式会社 for convenience. We remove the 株式会社, 有限会社, and 合同会社. As for the English names, we remove the designations like Co.,Ltd, Inc. and so on [1].

Remove special character Many Japanese company names contain special characters like ［・′＆．‵］. We remove these characters with regular expressions.

Normalization In this step, we covert the full-width (全角) characters to half-width characters if the company names contain latin characters or katakana. For example, we change Ｖｉｄｅｏｇｒａｐｈｙ　ＡＳＩＡ株式会社 to Videography ASIA 株式会社. And we further add the lower case version and upper case version (e.g. VIDEOGRAPHY ASIA 株式会社 and videography asia). As for the company names that contain lower case katakana (小文字), we convert them to the upper case katakana (大文字). For example, converting ザレジェンド to ザレジエンド.

Remove duplicates from aliases There are two resources that generate duplicate names. One is that the trade name can be duplicate (e.g. there are 4 companies have the same name "TIS 株式会社"). There are total 2,935,124 unique company names after removing the duplicates. Because one of these duplicates may contain English name or katakana name, we first perform the step 1 to add the names if they are available. Another reason is the alias generation process would produce duplicates. For example, the "TIS 株式会社" and "株式会社 TIS" are two different companies. But after the step 3 and step 5, they will both generate "TIS" and "tis". So we need to remove such duplicates from the generated aliases.

Dataset	Mainichi	BCCWJ
Sentences	3,027	1,364
Entities	4,664	1,704
Unique Entities	1,580	897

Table 2: Datasets statistics

Lexicon Type	Lexicon Name	Total Entities	Company Entities
Single Lexicon	JCL	8,418,872	8,418,872
	IPAdic	392,216	16,668
	NEologd	3,171,530	244,213
	Juman	751,185	9,608
Multiple Lexicon	IPAdic-NEologd	4,615,340	257,246
	IPAdic-NEologd-JCL	13,034,212	8,584,608

Table 3: Lexicons statistics

3 Experiment

In this section, we first introduce the datasets and lexicons, then evaluate contributions of lexicons from intrinsic perspective and extrinsic perspective [2].

3.1 Datasets and Lexicons

Data In our experiments, we use two Japanese annotated corpora, the Balanced Corpus of Contemporary Written Japanese (BCCWJ)[3], and Mainichi Newspaper Corpus[4]. According to the entity annotation scheme[5], these datasets contains multiple entity types. But we only extract the samples that contain the "Company" type. There are total 4,391 sentences from two datasets (see Table 2).

Lexicons Table 3 lists the statistics of different lexicons. The total entities is the total number of entities of a lexicon. The company entities mean the number of "Company" type entities. As for the Japanese Company Lexicon (JCL) lexicon, there are total 8,418,872 alias names, which has the most company names than other single lexicons. We compare JCL lexicon with other famous single Japanese lexicons, IPAdic, NEologd [3], and Juman. We also add the widely used multiple lexicon, mecab-ipadic-NEologd [4], which represented as IPAdic-NEologd. The multiple lexicon means the lexicon contains multiple single lexicons. We add JCL to IPAdic-NEologd for further comparison with IPAdic-NEologd, represented as IPAdic-NEologd-JCL. Beside the JCL, other lexicons contains other entity types ("Person", "Location", etc.), so the number of total entities are usually larger than the number of company entities.

[3] https://pj.ninjal.ac.jp/corpus$_{c}enter/bccwj/$

[4] http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html

[5] https://sites.google.com/site/extendednamedentity711/top 以下の階層の全リスト

Lexicon Type	Lexicon Name	Mainichi		BCCWJ	
		Count	Coverage	Count	Coverage
Single Lexicon	JCL	803	0.5082	485	0.5407
	IPAdic	726	0.4595	316	0.3523
	NEologd	416	0.2633	230	0.2564
	Juman	197	0.1247	133	0.1487
Multiple Lexicon	IPAdic-NEologd	836	0.5291	412	0.4593
	IPAdic-NEologd-JCL	1,078	0.6823	585	0.6522

Table 4: Intrinsic evaluation for different lexicons. The count means how many company names covered by a lexicon in a given dataset. The coverage score is the ratio of the covered company names.

3.2 Intrinsic Evaluation

The intrinsic evaluation is to evaluate a lexicon with its own intrinsic characteristics, the coverage [5]. In other words, we measure how many entities in the a dataset could be covered by a lexicon. The definition of coverage is below:

$$coverage = \frac{\sum \delta(e_i, L)}{\sum_{i=0}^{n} e_i}, \tag{1}$$

where n is the total number of company entities in a dataset, e_i is the i-th company entity in the dataset, and $\delta(e_i, L)$ will return 1 if e_i is in the lexicon L. So $\sum \delta(e_i, L)$ returns the total entity counts. If the coverage is higher, it means the lexicon cover more company entities in a dataset. For example, the coverage of the IPAdic lexicon for Mainichi dataset is 0.4595. Because IPAdic lexicon cover 726 entities in Mainichi dataset (totally 1,580 unique entities).

Table 4 list the coverage scores for different lexicons and datasets. We can see the JCL lexicon cover most company names (0.5082 in Mainichi and 0.5407 in BCCWJ) than other single lexicon. As for the multiple lexicons, JCL also boost the coverage score of IPAdic-NEologd (from 0.5291 to 0.6923 in Mainichi, from 0.4593 to 0.6522 in BCCWJ).

3.3 Extrinsic Evaluation

The extrinsic evaluation is to evaluate a lexicon with a downstream NLP task to see how much the lexicon information contributes to performance [6]. Here we choose the Named Entity Recognition (NER) task.

3.3.1 Experimental Setup

Annotation with lexicons In order to measure the contributions of different lexicons, we utilize the information contained in a lexicon for the training process. More specifically, we annotate the sentence in the character-based level if it contains company names. Then we save the annotated data as the IOB2 format. For example, assuming that we have a sentence “岩崎産業が運行する貨物フェリー”, and have “岩崎産業” and “岩崎” in the lexicon. Then we perform the matches in a greedy fashion by choosing the longest possible match in different lexicons, so the corresponding tags should be `["B-company", "I-company", "I-company", "I-company", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O"]`. Beside these tagged labels, we call the true labels as gold annotation. JCL contains many single character names and digital names, such as “風”, “1990”” and so on. We do not annotate them as company names in case of annotation errors.

Model We use two models for the training, the Conditional Random Field (CRF) model and the Bidirectional LSTM-CRF (Bi-LSTM-CRF) model. The CRF model is a traditional machine learning method that need manual feature engineering, and the Bi-LSTM-CRF model is a neural network based method that creates features automatically by the Bi-LSTM layer.

Dataset split For the CRF model, we split datasets with 70% training set and 30% test set (2,118 and 909 in Mainichi, 954 and 410 in BCCWJ). As for the Bi-LSTM-CRF model, we split datasets with 70% training set, 15% validation set and 15% test set (2,118, 404 and 405 in Mainichi, 954, 205 and 205 in BCCWJ). The labels of training set are annotated by different lexicons. The labels of training set are tagged labels by lexicons, and the labels of validation set and test are true labels.

Training setup For the CRF model, we use the L-BFGS based gradient descent to update the parameter. We set both c_1 (the coefficient for L1 regularization) and c_2 (the coefficient for L2 regularization) as 0.01. The max iterations is 100. As for the Bi-LSTM-CRF model, the embedding size of character is 128, hidden neuron size is 128, the batch size is 64, learning rate is 0.001, and the epoch is 15.

Metric We use the F_1 score as the metric for the NER task and train the model three times to take the average as the final score.

Lexicon Type	Lexicon Name	Mainichi		BCCWJ	
		CRF	Bi-LSTM-CRF	CRF	Bi-LSTM-CRF
Gold Annotation		0.9756	0.9692	0.9273	0.8924
Single Lexicon	JCL	0.8744	0.8894	0.8538	0.8591
	IPAdic	0.9383	0.9389	0.8682	0.8321
	NEologd	0.9136	0.9127	0.8456	0.8289
	Juman	0.9061	0.8949	0.8370	0.8105
Multiple Lexicon	IPAdic-NEologd	0.9261	0.9329	0.8657	0.8461
	IPAdic-NEologd-JCL	0.8761	0.8903	0.8551	0.8593

Table 5: Extrinsic evaluation for different lexicons with F_1 score. The gold annotation means the labels of training set are true labels.

3.3.2 Results

Table 5 shows the F_1 scores when using different lexicons. The gold annotation result means training the models with true labels, so the results are the upper bound for different lexicons. The result of CRF model are slightly better than the result of Bi-LSTM-CRF generally. The reason might be the limited data size. To make full use of neural network based model, it usually needs a large amount of data to learn the features automatically. But we only have 2,118 sentences in Mainichi and 954 sentences in BCCWJ for training, which limits the learning ability of Bi-LSTM-CRF model.

For the Mainichi dataset, other lexicons (espesilly the IPAdic, NEologd, and IPAdic-NEologd) perform better than JCL. This is because JCL contains many short form aliases, like “今日”” and “生命”, which will cause annotate error. This is very clear if we compare the F_1 score of IPAdic-NEologd and IPAdic-NEologd-JCL. Another reason is the dataset bias. Mainichi dataset contains data that mainly from news and the these news are most related to listing or famous companies. In other words, Mainichi dataset contains few long tail company names, which is hard to measure the true contribution of JCL for the long tail company detection. Because most of lexicons acquire data from web and news according to the frequencies, so their entities quality are higher than JCL, which contain less annotation error. For the BCCWJ dataset, the F_1 score of all lexicons are very close. The BCCWJ dataset contains more long tail company names than Mainichi dataset, so JCL performs well on both CRF and Bi-LSTM-CRF models.

4 Conclusion

In order to solve the long tail company name recognition problem and the variant forms recognition problem, we propose an alias generation process and build a high coverage lexicon for Japanese company name recognition. Because of the high coverage advantage, JCL can find long tail company names as much as possible, which makes JCL a good complement for other frequency-based lexicons in the company recognition field. Another advantage of JCL is easy to acquire and construct, which is suitable for coarse grained annotation to avoid human annotation cost. In the future work, we will redesign the annotation rule to decrease the annotation error.

References

[1] M. Loster, Z. Zuo, F. Naumann, O. Maspfuhl, and D. Thomas, “Improving company recognition from unstructured text by using dictionaries.,” in *EDBT*, 2017.

[2] S. Mohammad, C. Dunne, and B. Dorr, “Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus,” in *EMNLP*, 2009.

[3] T. H. Toshinori Sato and M. Okumura, “Operation of a word segmentation dictionary generation system called neologd (in japanese),” in *IPSJ-SIGNL*, 2016.

[4] T. H. Toshinori Sato and M. Okumura, “Implementation of a word segmentation dictionary called mecab-ipadic-neologd and study on how to use it effectively for information retrieval (in japanese),” in *Proceedings of the Twenty-third Annual Meeting of the Association for Natural Language Processing*, 2017.

[5] K. Labille, S. Gauch, and S. Alfarhood, “Creating domain-specific sentiment lexicons via text mining,” 2017.

[6] W. L. Hamilton, K. Clark, J. Leskovec, and D. Jurafsky, “Inducing domain-specific sentiment lexicons from unlabeled corpora,” in *EMNLP*, 2016.