

Introduction to Datasets

Zhouhao Wang

December 31, 2017

1 The Reuters News database

Here we are going to briefly introduce the Reuters News database in three aspects, including the basic information about the Thomson Reuters News agent, the structure of the news database we use, and the method to extract cross-lingual news text pairs from this database. This dataset is for sell on Thomson Reuters News' website. Please find more detail information on

<https://financial.thomsonreuters.com/en/products/infrastructure/trading-infrastructure/elektron-enterprise-data-management.html>

1.1 Introduction of the Thomson Reuters News

Thomson Reuters news ¹ is a world wide news agency located in London, England, and provides worldwide real-time breaking news and high-impact global multimedia content in multiple languages, especially focusing on the markets, business, politics, entertainment, technology, video and pictures. Although most of the reports are originally written in English in spite of the country or region they are related to, there are large proportions of them translated and edited into other languages including Chinese, Japanese and so on. These multi-lingual texts are expected to be highly potential resources for tasks related the multi-lingual natural languages processing. In this research, we use the database only related to the economics.

1.2 Database structures

The database we obtained has the structures shown in the Figure1. There are mainly 20 fields for each piece of news where what matters us in this research includes the field: HEADLINE_ALERT_TEXT, PNAC, TAKE_TEXT as well as LANGUAGE. The HEADLINE_ALERT_TEXT field contains the short titles, brief summarizations or alert messages for each piece of news, and we extract them as short text resources in the future experiments. TAKE_TEXT is the field for the content of stories, normally containing complete text of stories, comments and references, and we use them as source of long text. PNAC shorts for Primary News Access Code. According to the official guidance, it is the "Content Identifier that applies to all news messages relating to the same story", and we use this code as the search objects during building the cross-lingual word pairs. LANGUAGE indicates the language this report is written in and due to the requirements of this task, we only keep those texts written in English and Japanese based on this field. For more information about this database, you can refer to the official guidance THOMSON REUTERS NEWS FEED DIRECT. ².

1.3 Various forms of news text

The Thomson Reuters News database in economics contains not only the pure text news articles, but also includes charts, video/photo illustration as well as short market/stock reports.

¹Official websites of Thomson Reuter: <http://www.reuters.com/>

²More specific description of database guidance could be find on:
<http://share.thomsonreuters.com/assets/elektron/news-feed-direct-overview.pdf>

Field	Example
DATE	2014-12-01
TIME	00 05 12.906
UNIQUE_STORY_INDEX	20141201000512nL3N0TK0UA
EVENT_TYPE	STORY_TAKE_OVERWRITE
PNAC	nL3N0TK0UA
STORY_DATE_TIME	2014-12-01 00:05:12
TAKE_DATE_TIME	NaN
HEADLINE_ALERT_TEXT	香港デモ隊と警察が衝突、政府本部庁舎近く 少なくとも18人逮捕
ACCUMULATED_STORY_TEXT	NaN
TAKE_TEXT	〔香港 1日 ロイター〕 - 民主派の大規模デモが続く香港で1日未明、政府本部庁舎を包囲しようとした数千人のデモ隊に警官隊が警棒で殴りつけたり、催涙スプレーを噴射したりした。政府本部庁舎のある香港島の金鐘（アドミラルティ）では、デモを主導する学生団体が抗議行動を強化しようと呼び掛け、呼応した群衆が集まっていた。警察当局によると、衝突に絡んで少なくとも18人が逮捕された。多数の群衆は防護眼鏡などを身に付け、政府本部庁舎周辺地域からの撤退を拒否。「われわれは普通選挙を望んでいる」と叫び、警官隊と引き続き衝突している。※英文参照番号 [nL3N0TK0AQ]
PRODUCTS	RJN RS RSS JWM DNP
TOPICS	EMRG HK CN ASIA PIA NEWS1 VIO JFOR JLN DIP GEN...
RELATED_RICS	NaN
NAMED_ITEMS	NaN
HEADLINE_SUBTYPE	NaN
STORY_TYPE	S
TABULAR_FLAG	False
ATTRIBUTION	RTRS
LANGUAGE	JA

Figure 1: Example of a piece of data and its description

2 Extraction of cross-lingual news pairs

The Thomson Reuters provides news worldwide in multiple languages including English, Japanese and Chinese and so on. However, the original database we obtain does not offer the direct mapped pairs for us to extract cross-lingual news pairs directly. Fortunately, as the bold texts indicates in the Figure1, some of the Japanese reports provide reference codes in the TAKE_TEXT field linking to specific English reports, which has a form similar to standard PNAC fields. We conduct the following four procedures to retrieve the English-Japanese text pairs related to the same event:

1. Extract Japanese news containing the keyword (reference code)
2. Match and extract the possible reference code using the regular expression 'reference code' `\\([\\PNAC]+)\\]`, where the **PNAC** is the possible reference code extracted in the previous step.
3. Use the list of reference codes obtained previously as the queries searching for the available English news in the whole database.
4. Extract the corresponding HEADLINE_ALERT_TEXT field for Japanese and English news as cross-lingual title (short text) pairs and corresponding TAKE_TEXT as cross-lingual content (long text) pairs.

Following is an example of Japanese-English pairs that is successfully retrieved for new title. ollowing is an excerpt example of Japanese-English pairs for news contents corresponding to the

English	Japanese
UPDATE 2-Hyundai, Kia face fading growth as currency tides buoy Japan rivals	韓国の現代・起亜自、2014年世界販売台数は4%増見込む 03年以来の低成長

Figure 2: Example of SHORT cross-lingual pairs

previous news title. The cross-lingual potential resources in the Thomson Reuters News database are enormous. In practical, we successfully extract the more than 60,000 Japanese-English news pairs, including around 60,000 title pairs and around 60,000 content pairs for the complete year of 2014. As a future work, more cross-lingual pairs in other languages and years might be further exploited. The following experiments on this dataset mainly base on these cross-lingual pairs.

3 Preprocessing text

Now we'll introduce the specific procedures of pre-processing for both English and Japanese text which is applicable to both news title (i.e. short text) and news articles (i.e. long text). For general data/text mining tasks, pre-processing refers to processing prior to feeding data to the training models, as its prefix 'pre-' suggests. The purpose of preprocessing typically includes

1. removing noises including typos;
2. regularizing text formats for those text in different forms, so that they all obey the same format rules.
3. removing unrelated subjects such as http addresses, figures, special characters and so on.

Although preprocessing methods varies according to the specific text mining task, they usually share some common processing procedures including tokenization, tagging, lemmatization, also called morphological analysis, as well as general normalization.

English	Japanese
<p>* S. Korean firms see 2014 car sales at 7.86 mln, up 4 pct</p> <p>* Quick weakening of yen favours exports by Japanese peers</p> <p>* Shares skid with won at 5-year high vs dollar, yen</p> <p>(Add closing share prices, production figures, analyst comments)</p> <p>SEOUL, Jan 2 (Reuters) - The man who led South Korea's auto industry on a tear through the last decade said Hyundai Motor Co < 005380.KS > and Kia Motors Corp < 000270.KS > expect what will be their lowest annual sales growth since 2003 as the weak yen fires up Japanese rivals.</p> <p>In his annual New Year speech to staff on Thursday, 75-year-old group chairman Chung Mong-koo said sales at Hyundai and its smaller affiliate Kia will likely grow just 4 percent in 2014. Global competition is about to get tougher in an industry facing changing technology and an uncertain future, he warned.</p>	<p>〔ソウル 2日 ロイター〕 - 韓国の自動車メーカー、現代自動車 < 005380.KS > とその系列の起亜自動車 < 005380.KS > は、2014年の世界での販売台数について、両社合計で前年比4%増の786万台との目標を設定した。内訳は現代が490万台で、起亜が296万台。規制当局に2日提出した文書で判明した。</p> <p>4%という伸び率は、2.3%だった2003年の以来の低水準。円安を背景に、日本のメーカーの輸出攻勢が強まることを織り込んだ。</p> <p>現代自の鄭夢九・会長は、従業員への新年のメッセージで「世界経済が低成長時代に入るなか、企業間の競争は激化している」と述べた。</p> <p>現代・起亜自の世界販売台数は、ウォン安で輸出競争力が高まった2010年には24%増を記録した。2013年の販売台数は両社合わせて6%増の756万台で、2012年の8%増から伸びが鈍化した。</p> <p>※英文参照番号 [nL3N0K90F2]</p>

Figure 3: An example of LONG cross-lingual pair (Excerpt)