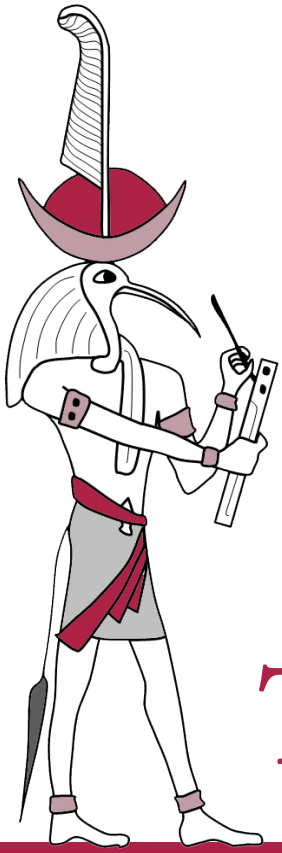
Terminologica



TOTh 2019

Terminologie & Ontologie : Théories et Applications

Presses Universitaires Savoie Mont Blanc

Terminologie & Ontologie : Théories et Applications

Actes de la conférence

TOTh 2019

Le Bourget du Lac – 6 & 7 juin 2019

Les ouvrages TOTh précédents sont disponibles sur le site du Comptoir des Presses d'Universités (www.lcdpu.fr) ou auprès de : contact@toth.condillac.org

Éditeur : Presses Universitaires Savoie Mont Blanc
27 rue Marcoz
BP 1104
73011 CHAMBÉRY CEDEX
www.univ-smb.fr

Réalisation : C. Brun, C. Roche
Collection « Terminologica »
ISBN : 978-2-919732-80-7
ISSN : 2607-5008
Dépôt légal : juillet 2020

Terminologie & Ontologie : Théories et Applications

Actes de la conférence

TOTh 2019

Le Bourget du Lac – 6 & 7 juin 2019

http://toth.condillac.org

avec le soutien de :
Université Savoie Mont Blanc
École d'ingénieurs Polytech Annecy Chambéry

Presses Universitaires Savoie Mont Blanc
Collection « Terminologica »

Comité scientifique

Président du Comité scientifique : Christophe Roche

Comité de pilotage

Rute Costa	Universidade Nova de Lisboa
Humbley John	Université Paris 7
Kockaert Hendrik	University of Leuven
Christophe Roche	Université Savoie Mont Blanc

Comité de programme 2019

Le comité de programme est constitué chaque année à partir du comité scientifique de TOTh en fonction des soumissions reçues. La composition du comité scientifique est accessible à l'adresse suivante : http://toth.condillac.org/committees

Guadelupe Aguado	Universidad Politécnica de Madrid – Spain
Amparo Alcina	Universitat Jaume I – Spain
Bruno Bachimont	Université Technologie de Compiègne – France
Jean-Paul Barthès	Université Technologie de Compiègne – France
Christopher Brewster	TNO – The Netherlands
Danielle Candel	CNRS, Université Paris Diderot – France
Sylviane Cardey	Université de Franche-Comté – France
Stéphane Chaudiron	Université de Lille 3 – France
Manuel Célio Conceição	Universidade do Algarve – Portugal
Rute Costa	Universidade NOVA de Lisboa – Portugal
Bruno Courbon	Université Laval – Canada
Lyne Da Sylva	Université de Montréal – Canada
Luc Damas	Université Savoie Mont-Blanc – France
Éric De La Clergery	INRIA – France
Dardo De Vecchi	Kedge Business School – France
Valérie Delavigne	Université Paris 3 – France
Sylvie Després	Université Paris 13 – France
Juan Carlos Diaz Vasquez	EAFIT University – Colombia
Hanne Erdman Thomsen	Copenhagen Business School – Denmark
Pamela Faber	Universidad de Granada – Spain
Christiane Fellbaum	Princeton University – USA
Iolanda Galanes	Universidade de Vigo – Spain
Christian Galinski	INFOTERM – Austria
François Gaudin	Université de Rouen – France
Teodora Ghiviriga	Alexandru Ioan Cuza University – Romania
Jean-Yves Gresser	ancien Directeur à la Banque de France – France
Ollivier Haemmerlé	Université de Toulouse – France
Gernot Hebenstreit	University of Graz – Austria
Amanda Hicks	University of Florida – USA
John Humbley	Université Paris 7 – France

Kyo Kageura	University of Tokyo – Japan
Heba Lecocq	INALCO – France
Hélène Ledouble	Université de Toulon – France
Patrick Leroyer	Aarhus University – Denmark
Georg Löckinger	University of Applied Sciences Upper Austria – Austria
António Lucas Soares	University of Porto, INESC – Portugal
Bénédicte Madinier	Dispositif d'enrichissement de la langue française – France
Candida Jaci de Sousa Melo	Universidade Federal do Rio Grande do Norte – Brazil
Jean-Guy Meunier	Université de Montréal – Canada
Christine Michaux	Université de Mons – Belgium
Fidelma Ní Ghallchobhair	Foras na Gaeilge, Irish-Language Body – Ireland
Henrik Nilsson	TNC – Sweden
Silvia Piccini	Italian National Research Council – Italy
Suzanne Pinson	Université Paris Dauphine - France
Marina Platonova	Riga Technical University – Latvia
Thierry Poibeau	CNRS Lattice – France
Maria Pozzi	el colegio de méxico – Mexico
Michele Prandi	Università degli Studi di Genova – Italy
Jean Quirion	Université d'Ottawa – Canada
Renato Reinau	Suva – Switzerland
Christophe Roche	Université Savoie Mont Blanc – France
Mathieu Roche	CIRAD – France
Laurent Romary	INRIA & HUB-ISDL – Germany
Micaela Rossi	Università degli studi di Genova – Italy
Bernadette Sharp	Staffordshire University – Great Britain
Marcus Spies	Universität München - Germany
Anne Theissen	Université de Strasbourg – France
Philippe Thoiron	Université Lyon 2 – France
Marc Van Campenhoudt	Université libre de Bruxelles – Belgium
Kara Warburton	City University of Hong Kong – China
Maria Teresa Zanola	Università Cattolica del Sacro Cuore – Italy
Fabio Massimo Zanzotto	University of Roma – Italy

Avant-propos

La Terminologie est une discipline scientifique à part entière qui puise à de nombreux domaines dont la linguistique, la théorie de la connaissance et la logique. Pour que cette diversité soit une richesse, il faut lui offrir un cadre approprié au sein duquel elle puisse s'exprimer et s'épanouir : c'est une des raisons d'être des Conférences TOTh créées en 2007. A ces conférences « mères » qui se tiennent chaque année à l'Université Savoie Mont-Blanc sont associées depuis 2011 les Journées d'étude TOTh dédiées à un thème plus spécifique organisées par une institution partenaire.

Dans ce contexte, la formation et la transmission des connaissances jouent un rôle essentiel. La *Formation TOTh* précédant la Conférence se déroule sur deux années consécutives dédiées pour l'une à la dimension linguistique et pour l'autre à la dimension conceptuelle de la terminologie, deux dimensions étroitement liées.

A la présentation de travaux sélectionnés par un Comité de programme international, la *Conférence TOTh* inclut une *Conférence invitée* et, selon les années, une *Disputatio*. La première, donnée par une personnalité reconnue dans son domaine vise l'ouverture à d'autres approches de la langue et de la connaissance. La seconde, à travers une lecture commentée effectuée par un membre du comité scientifique, renoue avec une forme d'enseignement et de recherche héritée de la scolastique.

Christian Galinski de Infoterm, a ouvert la conférence sur le sujet de « *The emergence of terminology science and terminological activities* ».

Cette année, comme en 2018, nous n'avons pas inclus de *Disputatio* par manque de temps. En effet, pour la première fois, TOTh a accueilli une session satellite, en parallèle avec la conférence, sur le thème de « Terminology and Text Mining » en lien direct avec les thèmes de TOTh. Nous avons également dédié une session de la conférence au projet Européen ELEXIS.

Les 29 communications et les 3 posters ont permis d'aborder de nombreux sujets tant théoriques que pratiques, autant d'exemples de la diversité et de la richesse de notre discipline. Je vous invite à découvrir à travers ces actes les 24 interventions qui ont donné lieu à publication.

Avant de vous souhaiter bonne lecture, j'aimerais terminer en remerciant tous les participants pour la richesse des débats et des moments partagés.

Christophe Roche
Président du comité scientifique

Sommaire

Chinese Word Segmentation with External Lexicons on Patent Claims

Yixuan Li*, Kim Gerdes**

*LPP (CNRS), Université Sorbonne Nouvelle - Paris 3,
19 Rue des Bernardins, 75005 Paris,
yixuan.li@sorbonne-nouvelle.fr
**Almanach (INRIA), 2 Rue Simone IFF, 75012 Paris,
kim@gerdes.fr

Abstract. This paper aims to compare the performance of different Chinese word segmenters in specialized technological domains as well as to evaluate the contribution of an external lexicon to their improvement. As we are interested in patent texts whose automatic analysis is of economic and scientific importance, we attempt to tackle the hardest source text for terminology extraction in terms of language and genre : Chinese patent claims. Some previous work on Chinese segmentation adaptation to patents are based on training a new model or using predefined term dictionaries, and none focuses on the adaptability of existing state-of-art segmenters. Our approach uses raw textual patent claim data, both supervised and unsupervised state-of-the-art word segmenters, technological dictionaries, entropy measures for wordhood detection, and most importantly an automatic approach to build large reliable lexicons. We show how much each resource contributes to finding the best segmentation.

1. Introduction

This paper attempts to find an optimal solution for the Chinese patent segmentation by combing different pre-trained models with accessible external resources and avoiding any costly annotation. We are most interested in patent claims, whose automatic analysis including data mining and terminology extraction is of significant economic and scientific importance, especially in the context of rapid accumulation of Chinese patent applications since a decade.

Patents in all languages are notoriously rich in new terminology, and inside a patent, the claims, the legally binding part of the patent, contain an even denser terminology than the patent description or the patent abstract. Moreover, the obligation to express each claim in one single sentence makes the structure very different from standard language and particularly hard to analyze. Using syntactic analyzers to process the raw textual patent claim data has been proven helpful in the task of terminology extraction from patents (Yang and Soo 2012). Chinese, however, is a *scriptua continua*, and therefore in a general NLP pipeline, before syntactic parsing and all other kinds of downstream tasks, the chain of characters should be cut into tokens. This is not a problem for common texts with f-scores usually beyond 97% (Zhao and al. 2017). However, if we apply an out-of-the-shelf Chinese word segmenter on patent claims, we have a high percentage of words the segmenter has never seen, so-called Out-of-Vocabulary (OOV) words that considerably degrade the results. All the more tricky but also fascinating from a term extraction perspective is the high percentage of the OOV that are newly created terms, which have not been recorded in any dictionary yet.

Different from many previous works on domain adaptation of Chinese word segmentation, instead of training a new system on abundant annotated data hard to be updated and thus unsuitable for such a domain like patent application changing with each passing day, our work concentrates on evaluating the adaptability and extensibility of general segmenters. The successful adaptation of general segmenters can avoid time-consuming manual corpus labeling works to train a specialized model on specific domain every time. In our experiments, we also verify several hypotheses derived from the data: (1) covering the OOV terms with massive custom dictionaries may help to improve the results; (2) the quality of segmentation may vary between IPC classes; (3) the unsupervised method may have a better performance on the domain-specialized segmentation.

After the brief presentation of linguistic specificities of Chinese language and patent claims, we analyze current difficulties in the word segmentation on Chinese patent claims in section 2. Then, we introduce in section 3 our methods of construction of the external lexicons that are used later in experiments, and in section 4 our annotation framework on test dataset. In the last section, we show the final results and an ablation study on the improvement.

2. Chinese Word Segmentation in specialized domains

As a *scriptua continua* and an isolating language, unlike many alphabetic languages such as French and English, Chinese does not have naturally recognizable linguistic units within written sentences, namely "words". Instead, the sentence in Chinese is a continuous series of characters containing neither white space nor any kind of distinguishable word boundary markers. Only based on this fact, can we understand the long-lasting debate around *wordhood* in Chinese.

In this section, we discuss the wordhood in Chinese from both a linguistic and a technological perspective and show with examples where reside problems of adaptation of segmenters to specialized technological domains.

2.1. Wordhood in Chinese

	咖啡 **ka-fei**	一个 **yi-ge**	小朋友们 **xiao-peng-you-men**
gloss	(transliterated)	one -quantifier	little-friend-friend-plural
meaning	*"coffee"*	*"one" "a(n)"*	*"children"*
GB	咖啡	一 个	小朋友 们
UD	咖啡	一 个	小朋友们
LTP/ THU/ JIE	咖啡	一个 / 一 个	小朋友们 / 小 朋友 们 / …

Tab. 1 – *Examples of the incoherence in Chinese word segmentation between different corpora and standards. GB is the GuoBiao standards[1], UD stands for the Universal Dependencies treebanks[2], and LTP/THU/JIE stand for three state-of-the-art segmenters[3] that are also used later in our*

1 GB/T 13715-1992 *Contemporary Chinese language word segmentation specification for information processing* (《信息处理用现代汉语分词规范》) http://openstd.samr.gov.cn/bzgk/gb/newGbInfo?hcno=B48FFFB924DF90488FEBCB89B91C8869

2 https://universaldependencies.org

3 LTP stands for the segmenter in Language Technology Platform (LTP) (https://www.ltp-cloud.com/), a set of online learning toolkits developed by the Harbin Institute of Technology.

experiments (see section 5).

From the indiscriminate application of western linguistic notion to the most radical opposite that "Chinese does not have words, but instead has characters" (Hoosain 1992; Xu 1997; Packard 2000), currently no common agreement on the definition of "words" in Chinese has been reached. With the rapid development of information technology, the information processing on Chinese language faces a dilemma: While most of the popular tools that are originally developed for western languages require word breaking, the heavy manual process and low inter-annotator agreement make it hard to provide high-quality input corpora to downstream tasks.

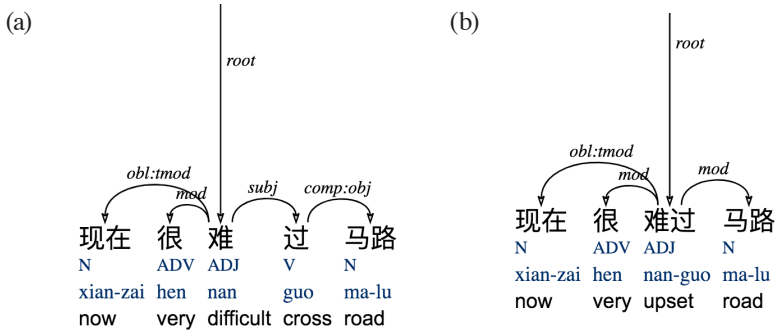
In Section 1.1, we investigate the incoherence of existing segmentation criteria: While terms such as "咖啡 ka-fei" have only one possible segmentation, "一个 yi-ge" and "小朋友们 xiao-peng-you-men" are more ambiguous cases in which all of these alternatives should be considered correct. In fact, without absolute standards, the required segmentation largely depends on these downstream tasks. Therefore, the segmentation standards lack practical meaning when the final application objective has not been fully considered.

2.2. Chinese word segmentation for patent claims

The Chinese Word Segmentation (CSW) is very often considered as the first step of various NLP tasks on Chinese, and thus it has unavoidable effects on all kinds of downstream tasks. Comparing to general texts, with their complexity in style and the high percentage of OOV, the patent claims can suffer from even greater noise introduced in segmentation (Fig. 1).

THU stands for the segmenter in THU Lexical Analyzer for Chinese (THULAC) (http://thulac.thunlp.org/), a Chinese NLP toolkit released by Thuhua University.

JIE stands for jiebe (Chinese for "to stutter") segmenter, an individual-developed Chinese Word Segmenter (https://github.com/fxsjy/jieba).



"Now it's difficult to cross the road."

Two main streams of segmentation algorithms are the lexicon-based (Chen et al. 1999 ; Nie, Jin and Hannan 1994) and statistics-based methods. At present most of the popular Chinese segmenters belong to the later that regard the segmentation task as a continuous sequence labeling problem (Ng and Low 2004 ; Low and al. 2005).

Research by Huang (2006) demonstrates that the segmentation errors caused by OOV are in general five times more important than in other cases. Theoretically, once the lexicon has fully covered the vocabulary of the dataset to segment, there should remain few errors concerning only the true ambiguity in original sentences. It is then a natural thought to use domain-specialized lexicons to improve the segmenter's performance on terms never seen in its training dataset. To construct the external dictionary, the simplest and most frequently used resource is the technological dictionaries (Zhao et al. 2010). However, the available dictionaries are unbalanced in terms of domains and often do not include the latest technologies (Rong 2015). Also, regarding the huge quantity and the considerable update speed of terminology in patent applications, a better way is to extract term lists directly from newly published patent applications and research papers carrying the most recent technological terms, and combine them into the pre-trained model.

In addition to their lexical specificities on OOV terms, at sentence level, patent claims are semi-structured texts with legalese expressions and extremely long sentences very often containing more than 100 characters compared to 20 or 30 characters in general texts. The unusual length of claim sentences, which increases the computing difficulty, is another cause of the low segmentation accuracy on patent claims.

published patent applications and research papers carrying the most recent technological terms, and combine them into the pre-trained model.

In addition to their lexical specificities on OOV terms, at sentence level, patent claims are semi-structured texts with legalese expressions and extremely long sentences very often containing more than 100 characters compared to 20 or 30 characters in general texts. The unusual length of claim sentences, which increases the computing difficulty, is another cause of the low segmentation accuracy on patent claims.

> 3 . 如 权利 要求 2 所述 的 一种 荧光 定量 PCR 检测 鼠疫 耶尔森 氏菌 的 方法 ， 其 特征 在于 ：荧光 定量 PCR 标准 曲线 采用 以 下 步骤 制得 ： 取 不同 DNA 载量 的 质粒 参考 品 各 2 μ l ， 按 上述 荧光 定量 PCR 的 反应 体系 及 反应 程序 在 实施 荧光 定量 PCR 仪|上 进行 扩增 ； 反应 结束 后 根据 得到 的 各个 浓度 的 循 环 阈值 C (t) ， 采用 计算机 自动 绘制 荧光 定量 PCR 标准 曲线 。
>
> *3. The method of **claim** 2, a fluorescent quantitative PCR detection of **Yersinia** pestis claim, wherein: quantitative PCR standard curve was prepared using the following steps: take different amounts of plasmid DNA contained in each of the two reference products μ l, the above-described quantitative PCR reaction system and the reaction procedure in an amplification on the PCR system; after completion of the reaction according to the cyclic threshold value C for each concentration obtained in (t), using the computer to automatically draw the quantitative PCR standard curve.*

The example above is one claim from an actual patent application on Yersinia pestis detecting[5], segmented by the Jieba segmenter (See section 5). The whole claim sentence in Chinese is composed of 137 characters (more than 88 words in English). The underlined character sequences in the example above are where the segmentation errors are found: Even "**权利要求** qian-li-yao-qiu", one of the most frequent legalese term in patent texts meaning "claim" is wrongly segmented into "**权利** (right)" and "**要求** (requirement)". The medical term "**耶尔森氏菌** (Yersinia)" and "PCR**仪 上** (on the polymerase chain reactor)" are other two segmentation errors in the example sentence: the former is segmented at a position where it should not be, the latter, in contrast, is not correctly segmented where it should be. Typical segmentation errors in this example sentence reveal the difficulty of patent claim processing.

5 Yersinia pestis is the plague bacterium. Patent CN102146467A applied by the Zhejiang Interna-tional Tourism Healthcare Center in 2001.

Construction of lexicons

	Domain	WIKI	CNKI	ELeVe
A	Human Necessities	34 181	9 965	4 157
B	Performing Operations ; Transporting	25 681	9 182	4 226
C	Chemistry ; Metallurgy	27 379	6 820	3 030
D	Textiles ; Paper	10 747	4 841	2 705
E	Fixed Constructions	14 421	8 751	3 613
F	Mechanical Engineering ; Lighting ; Heating ; Weapons ; Blasting	14 308	2 929	3 814
G	Physics	36 281	10 013	3 524
H	Electricity	23 823	4 861	3 824

Tab. 2 – *The International Patent Classification (IPC), established by the Strasbourg Agreement 1971, provides a hierarchical system of language independent symbols for the classification of patents and utility models according to the different areas of technology to which they pertain. In column WIKI, CNKI and ELeVe, we show the total number of terms extracted from the three resources.*

According to WIPO, patents are classified by the IPC standard in technological domains and subdomains. The fundamental distinction is the 8 classes from A to H (Tab. 2). We developed for each IPC classes a custom dictionary, which concatenates three distinct resources into one big word list : We took 1. all page titles from Chinese Wikipedia, 2. keywords used for classifying the academic papers on CNKI.net, and 3. a list of highly autonomous words that was produced as follows : The ELeVe algorithm (Magistry and Sagot 2012) analyzes raw textual data and computes the entropy after each character, i.e. the degree of freedom that the preceding characters offer for the next character.

3. Wikipedia page titles

As the largest online encyclopedia, Wikipedia allows free download of its dump dataset[6]. The zip file downloaded includes the titles of all Chinese Wiki pages (one page title per line and 5 420 881 terms in total in June 2019).

We finally extracted 18 681 Wikipedia page titles in total, divided into 8 domains – note that many titles are long and do not bother the segmentation as long as the precise string is not encountered, and to keep the word lists shorter, we kept only strings that actually appear in the patent domain data of each IPC class by overlapping it with our raw patent application texts and conserving only the termes appearing in the corpus.

3.1. CNKI document keywords

CNKI.net (China National Knowledge Infrastructure, 中国知网) is a Chinese key national information construction project to build and maintain a comprehensive Integrated Knowledge Resources System, including journals, doctoral dissertations, masters' theses, proceedings, newspapers, yearbooks, statistical yearbooks, ebooks, patents, standards, etc. For each of the 8 IPC domains, we construct an article keywords collection by web crawling academical article pages on the site. The classification is based on domain tags on CNKI.net. The obtained keywords are then concatenated to the former Wiki title lists. And again by filtering terms never seen in the patent corpus we reduce the term lists to a reasonable size. The numbers of terms extracted from Wikipedia and from CNKI.net is shown in Tab. 2.

3.2. Lists of term candidates produced by ELeVe

The unsupervised language-independent tokenizer - ELeVe (Magistry and Sagot 2012) is based on the computation of autonomy scores of multi-character terms by measuring the entropy between the characters. The idea of entropy-based segmentation is that a high entropy point is a good potential position for word segmentation, in particular, if the analysis is done bidirectionally. We make use of the ELeVe not only as a segmentation tool (section 5) but also as a resource to produce lists of highly probable terms.

For each IPC class, the segmenter thus provides us with a list of potential words that can be sorted by their degree of autonomy. However, in order to

6 https://dumps.wikimedia.org/zhwiki/

establish a reliable term lists from this raw list, which contains both good terms and bed terms (strings of characters that are not words), we built an automatic perceptron binary classifier model trained on positive and negative examples, selected from the top and the bottom of the list and filtered manually.

From the list of potential words in each IPC domain, we manually selected 2000 positive examples and 2000 negative examples. We also trained a Word-2Vec representation (Mikolov et al. 2013) on our raw corpus by using individual characters as neighbors of the potential words. This allowed us to add to each potential word the vector representations of its 10 closest potential words in terms of their distribution, something that resembles a list of synonyms. For each of these "synonyms", we also provided their degree of autonomy. We gave these 4000 words with different feature combinations of the word itself and for their potential synonyms to the perceptron, which thus trained a Chinese term discriminator model in order to distinguish good from bad term candidates with its best score close to 95%. Features given in training include:

- VecSyn, the vector representations of the 10 most close synonyms of the term
- VecOwn, the vector representation of the term itself
- DistSyn, the distances between the term and its 10 most close synonyms
- AutoSyn, the autonomy scores of the 10 most close synonyms

By means of an ablation study, we obtain the contribution of each feature, shown in Tab. 3. In F-measure results, the perceptron provided only with the vector representation of the term itself and the distances between the term and its 10 most close synonyms has the best accuracy. And the feature that contributes the most is the vector representation of the term itself (accuracy of 0.92 when given only the vector representation of the term itself).

VecSyn	VecOwn	DistSyn	AutoSyn	F-measure
Y	Y	Y	N	0.8892
Y	Y	N	N	0.9304
Y	N	Y	N	0.8636
Y	N	N	N	0.8793
N	N	Y	N	0.8096
N	Y	Y	N	0.9446
N	Y	N	N	0.9219
Y	Y	Y	Y	0.8991

Tab. 3 – *Combinations of different features given to the perceptron have an influence on accuracy. The best result corresponds to the model trained only with VecOwn and DistSyn.*

The high accuracy allows us to dress a large and reliable enough list of potential terms for each of the eight main IPC class with comparatively little manual efforts.

4. Annotation of raw patent claims

With no segmentation evaluation dataset available on Chinese patent claims, we construct our gold test set by randomly selecting a list of 100 lines of claims of each IPC class and segmenting them into segmentation units based on the Guobiao standards[7] and syntactic tests for those terms not found. In his work on customizable segmentation, Wu (2003) distinguishes five types of morphologically derived words in Chinese: 1. Reduplication, 2. Affixation, 3. Directional and resultative compounding, 4. Merging and splitting, 5. Named entities and factoids. We largely followed his classification in our annotation. And since word segmentation can be finer or coarser grained, and we want to compare segmentations that might differ in the granularity of their segmentation, our segmentation markers are annotated with unique symbols of the type of segmentation (Tab. 4): 1.completely syntactically free segmentation units, 2.multi-character expressions (possibly with ambiguous

7 GB/T 13715-1992 *Contemporary Chinese language word segmentation specification for information processing* (《信息处理用现代汉语分词规范》) http://openstd.samr.gov.cn/bzgk/gb/newGbInfo?hcno=B48FFFB924DF90488FEBCB89B91C8869

borders), 3.directional or resultative compounds, 4.classifiers, and 5.morphological affixes.

Marker	Type of cuts	Examples	Translation
space	syntactic unit boundaries	导通 和 断开	conducted and blocked
[]	multi-character expressions	[[中央] [处理器]]	CPU / Central Processor
()	directional/ resultative compounds	(设)有 (接)入	(set up) have_ AUX (link) in/into
|	classifiers	一|种	one | kind
{}	merging/splitting duplication etc. (i.e. idiom)	-	-
_	affixes	超_导体	super_conductor

Tab. 4 – *Distinct symbols used as segmentation markers to indicate six different types of cuts within a sentence. In our patent corpus, we barely found structures in the fifth group.*

We use unique symbols to distinguish six types of possible cuts (Tab. 4). Within all listed types, syntactic unit boundaries indicate the position between two independent syntactic units that without any doubt should be segmented; multi-character expressions markers annotate the inner structure of long terms, mostly noun phrases, which can be segmented on different granularity level, e.g. the term 中央处理器 (zhong-yang-chu-li-qi) can be annotated as a single unit meaning CPU while with a lower granularity we segment it into 中央 (zhong-yang, *"central"*) and 处理器 (chu-li-qi, *"processor"*); the analysis of directional/resultative compounds is reserved to serial verb (or auxiliary) constructions in which the second component denotes some sort of direction or result of the first component; classifiers or measure words are normally necessary between the numeral and the noun in Chinese (e.g. speakers say "one person" or "this person" in English, but "一个人 yi-ge-ren" or "这个人 zhe-ge-ren", here "个 ge" is the classifier); the fifth group is composed of other specific structures in Wu (2003), however, we barely found these structures in patent claims; and the last group is morphological affixes. We marked affixes in our test set only to distinguish them from independent

syntactic units and always regard them as inseparable parts of single words, that is to say, affixes and their radical are never segmented.

5. Experiments and results

We created a corpus of 347 950 claims from the SIPO patent application dataset from November 2017 to April 2018, which, after classification into eight classes, we processed to keep only the patent claims and stripped off all non-Chinese characters. Those characters, including Latin letters, Arabic numbers, punctuation, and all other symbols, are replaced by distinct separation symbols. Moreover, patent claims are highly standardized with specific legalese expressions such as the Chinese equivalents of "we claim", "disclosed is", "the composition of claim 1, wherein". Their number is limited, and they are of no interest for terminology extraction, and we also replaced them with unique symbols as placeholders.

The three supervised segmenters used in the paper are 1. Thulac (Sun et al. 2016), 2. pyltp (Che et al. 2010) and 3. Jieba (https://github.com/fxsjy/jieba). All three of them are state-of-the-art word segmenters frequently used as pre-processors in NLP research projects (e.g. Peng and Dredze 2015, Lizhen et al. 2014, Peng et al. 2017). And all of them accept a list of external terms as a custom dictionary, although they give different priority to the provided terms according to their algorithm.

The experiment results show that giving external word lists to the segmenter does improve the segmentation accuracy (Fig. 2), except for the Jieba segmenter, where we observe no improvements with the dictionary. For the other segmenters, we see that the larger the list, the better the accuracy. In all cases, the externally sourced Wikipedia and CNKI lists give better results than the ELeVe word list. However, the combination of all three lists gives the best results in general. This finding supports the first hypothesis that covering the OOV terms with massive custom dictionaries may help to improve the results.

On the other hand, the unsupervised method does not show a better performance in our experiments compared to supervised ones (Fig. 3). While the LTP and Jieba have no significant gap in segmentation accuracy, the ELeVe is always about 0.2 behind other supervised systems on all granularity levels. But it should be noted that the limitation of memory may prevent ELeVe from taking advantage of its learning ability on enormous data.

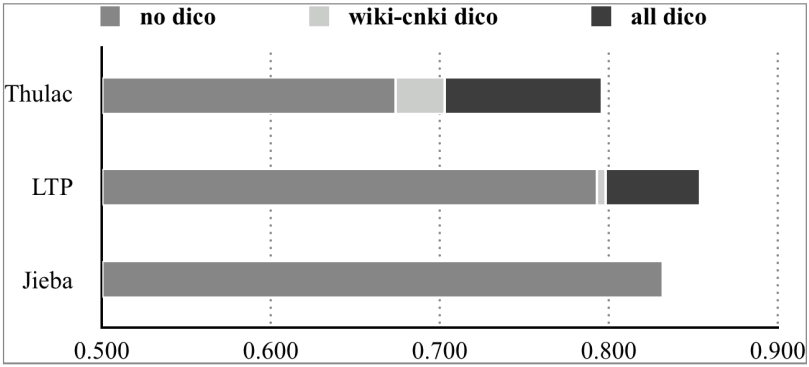


FIG. 2 – *Contribution of external lexicon on supervised segmentation accuracy. The accuracy in this experiment considers all possible cuts as correct segmentation.*

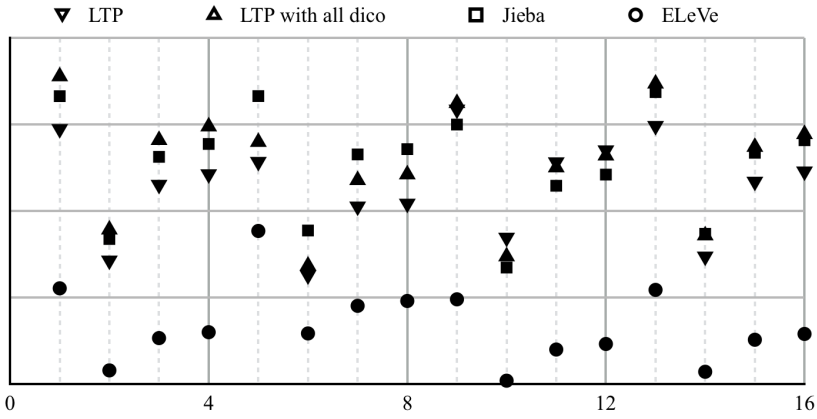


FIG. 3 – *Segmentation accuracy with different segmentation strategies and on different granularity levels. The horizontal axis present the accuracy of segmentation systems. On the vertical axis 1-4 are where we segment only syntactic unit boundaries (spaces); 5-8 segment also all directional/ resultative compounds in addition to all syntactic unit boundaries ; 9-12*

segment also classifiers in addition to all syntactic unit boundaries ; and 13-16 segment all three types above. Multi-character expressions are also segmented according to their granularity level (in each group the granularity level decreases with the growth of the number). For example, the triangle from top left is the accuracy of LTP segmenter with all dictionaries while in the gold file only syntactic unit boundaries (but not spaces inside directional/resultative compounds nor classifiers) are considered as segmentation boundaries as fine the granularity as possible.

In addition, we also observe differences in accuracy between IPC classes (Fig. 4), and as expected these gaps can be reduced with word lists. To investigate if the unsupervised ELeVe segmenter have better performance on larger training datasets, we use the white lines to present the size of dataset for each IPC class. The graph shows no obvious correlation between the size and the accuracy of ELeVe.

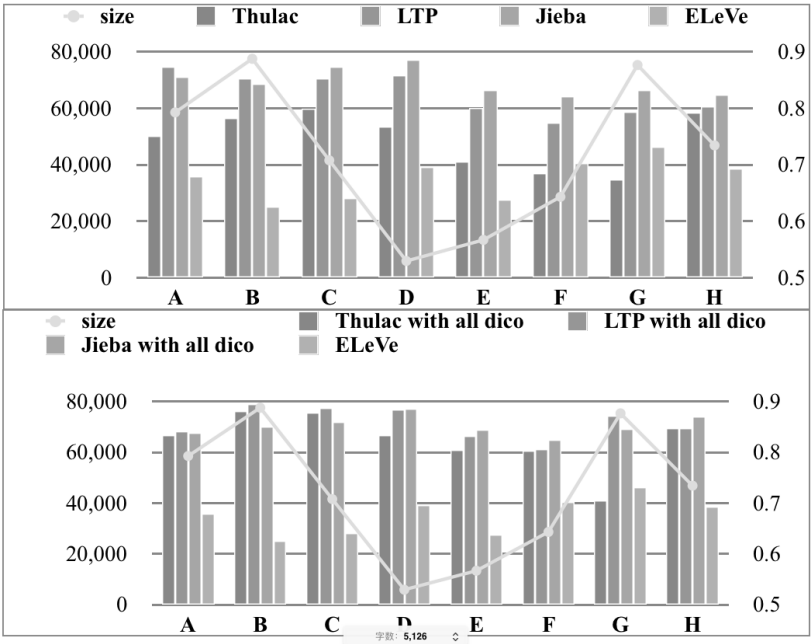


Fig. 4 – *Results of different segments on different IPC classes. As we were interested in the correlation between the performance of segmenters and the size of the corpus, we draw a white line indicating the size of each class.*

The accuracy evaluation on our test set is astonishingly complex but allows nonetheless to draw a few clear conclusions : Accuracy varies widely between 55 % to nearly 90 %, when we vary our parameters (the segmenter, the size of the vocabulary list, the source of the vocabulary list, the patent domain, the size of the required segments, and the types of the required word segmentation).

6. Conclusion

To sum up, the main contributions of the paper are fourfold : Firstly, this is the first work to compare systematically the performance of current segmenters on Chinese patent texts, especially the contribution of the coverage of unknown terms with the help of custom dictionaries extracted from different resources and of different size. Secondly, this is also the first study working on giving a segmentation gold standards to patent claims. Thirdly, the variability of the results shows that an evaluation that does not allow for different granularities of word segmentation cannot compare segmenters in a meaningful way. Lastly, we show how state-of-the-art machine learning techniques can supplement and enhance the extraction of large dictionaries even without a prior word segmentation.

We plan to test these methods again with contextual embeddings, which provide important precision gains on many NLP tasks. Another path is to overcome preliminary segmentation altogether by parsing technological texts with a model that has been trained on character-segmented treebanks. The first results of this method has been presented by Dong et al (2019).

Acknowledgements. We give our gratitude to the China Scholarship Council (CSC) for their financial support.

References

Che, Wanxiang, Zhenghua Li, and Ting Liu. 2010. “Ltp : A chinese language technology platform.” In Proceedings of the 23rd *International Conference on Computational Linguistics* : Demonstrations, pp. 13-16.

Dong, Chuanming, Yixuan Li, and Kim Gerdes. “Character-level Annotation for Chinese Surface-Syntactic Universal Dependencies.” 2019.

Du, Liping, Xiaoge Li, Gen Yu, Chunli Liu, and Rui Liu. 2016. “New word detection based on an improved PMI algorithm for enhancing segmentation

system." *Acta scientiarum naturalium universitatis pekinensis* 52, no. 1 : 35-40.

Gerdes, Kim. "Collaborative dependency annotation." In Proceedings of the second international conference on dependency linguistics (DepLing 2013), pp. 88-97. 2013.

Hoosain, Rumjahn. 1992. "Psychological Reality of the Word in Chinese." In H. C. Chen, & O. J. L. Tzeng (Eds.), *Language Processing in Chinese*, pp. 111-130.

Huang Changning, Zhao Hai. 2007. "Chinese Word Segmentation : A Decade Review." *Journal Of Chinese Information Processing*, 21(3): 8-19.

Liu, Lizhen, Song Wei, Wang Hanshi, Li Chuchu, Lu Jingli. 2014. "A novel feature-based method for sentiment analysis of Chinese product reviews." *China communications*, 11(3), 154-164.

Liu, Yuan, Q. K. Tan, and Xukun Shen. 1994. "Contemporary Chinese Language Word Segmentation Specification for Information Processing and Automatic Word Segmentation Methods."

Low, Jin Kiat, Hwee Tou Ng, and Wenyuan Guo. 2005. "A maximum entropy approach to Chinese word segmentation." In Proceedings of the SIGHAN Workshop on Chinese Language Processing, pp. 448–455.

Lu, Bin, and Benjamin K. Tsou. 2009. "Towards bilingual term extraction in comparable patents." In Proceedings of the 23rd *Pacific Asia Conference on Language, Information and Computation*, Volume 2, vol. 2.

Magistry, Pierre, and Benoît Sagot. 2012. "Unsupervized word segmentation : the case for mandarin chinese." In Proceedings of the 50th Annual Meeting of the *Association for Computational Linguistics* : Short Papers-Volume 2, pp. 383-387.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. "Distributed representations of words and phrases and their compositionality." In *Advances in neural information processing systems*, pp. 3111-3119.

Ng, Hwee Tou and Jin Kiat Low. 2004. "Chinese part-of-speech tagging : One-at-a-time or all-at-once ? word-based or character-based ?" In Conference on Empirical Methods in Natural Language Processing, pp. 277–284.

Packard, Jerome. 2000. *The Morphology of Chinese : A Linguistic and Cognitive Approach*. Cambridge University Press.

Peng, Haiyun, Erik Cambria, & Hussain, A. 2017. "A review of sentiment analysis research in Chinese language." *Cognitive Computation*, 9(4), 423-435.

Peng, Nanyun, Mark Dredze. 2015. "Named entity recognition for chinese social media with jointly trained embeddings." In Proceedings of the 2015 *Conference on Empirical Methods in Natural Language Processing* (pp. 548-554).

Song, Lifeng. 2011. "Research on Chinese Word Segmentation Algorithm for Patent Documents." *Straits Science* 7 : 9-11.

Sun, Maosong, Xinxiong Chen, Kaixu Zhang, Zhipeng Guo, and Zhiyuan Liu. 2016. "Thulac : An efficient lexical analyzer for chinese." *Technical Report*.

Wu, Andi. 2003. "Customizable segmentation of morphologically derived words in Chinese." *International Journal of Computational Linguistics & Chinese Language Processing*, Volume 8, Number 1, February 2003 : Special Issue on Word Formation and Chinese Language Processing 8, no. 1 : 1-28.

Xu, Yi, Alvin Meyer Liberman, and Douglas H. Whalen. 1997. "On the immediacy of phonetic perception." *Psychol. Sci.*, 8, 358–362 . Zanone, P. G. and Kelso, J. A.

Xun, Endong, and Li Cheng. 2009. "Applying terminology definition pattern and multiple features to identify technical new term and its definition." *Journal of Computer Research and Development* 46, no. 1 : 62-69.

Yang, Shih-Yao, and Von-Wun Soo. 2012. "Extract conceptual graphs from plain texts in patent claims." *Engineering Applications of Artificial Intelligence* 25, no. 4 : 874-887.

Yue, Jinyuan, Xu Jin'an, and Zhang Yujie. 2013. "Chinese word segmentation for patent documents." *Acta Scientiarum Naturalium Universitatis Pekinensis* 49, no. 1 : 159-164.

Zhai, Dongsheng, and Ma Wenshan. 2011. "Research the Algorithm of Chinese Patent Claims Segmentation." *Journal of Intelligence* 30, no. 11 : 152-155.

Zhang, Guiping, Liu Dongsheng, Yin Baosheng, et al. 2010. "Research on Chinese Word Segmentation for Patent Documents." *Journal of Chinese Information Processing*, 24(3): 112-116.

Zhang, Rong. 2015. *Terminology and Information Processing*. China Social Sciences Press.

Zhao, Hai, Huang Chang-Ning, Li Mu, and Lu Bao-Liang. 2010. "A Unified Character-based Tagging Framework for Chinese Word Segmentation." ACM Transactions on Asian Language Information Processing, 9(2), pp. 1–32.

Zhao Hai, Cai Deng, Huang Changning, Kit Chunyu. 2017. "Chinese Word Segmentation, a decade review (2007-2017)." The Frontier of Empirical

and Corpus Linguistics, Chunyu Kit and Meijun Liu ed., *China Social Sciences Press*.

Résumé

Cet article a pour objectif de comparer les performances de différents segmenteurs chinois dans des domaines technologiques spécialisés, ainsi que d'évaluer la contribution du lexique externe à leur amélioration. Nous nous intéressons aux textes de brevets dont l'analyse automatique revêt une importance économique et scientifique croissante, et nous tentons de nous attaquer à la source textuelle le plus difficile pour l'extraction de terminologie en termes de langue et de genre : les revendications de brevet chinois. Certains travaux antérieurs sur l'extraction de la terminologie des brevets chinois reposent sur entraînement d'un nouveau modèle ou utilisent des dictionnaires de termes prédéfinis, et aucun ne se concentre sur l'adaptabilité des segmenteurs état-de-l'art existants. Notre approche consiste à utiliser des données textuelles brutes des revendications de brevet, des segmenteurs supervisés et non-supervisés, des dictionnaires technologiques et des mesures d'entropie pour la détection de mots, et surtout une approche automatique permettant de construire de larges lexiques fiables. Nous montrons dans quelle mesure chaque ressource contribue à améliorer la segmentation adaptée.