The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence
Knowledge Extraction from Text: Technical Report WS-16-10

Coupled Semi-Supervised Learning for Chinese Knowledge Extraction

Leeheng Ma, Yi-Ting Tsao, Yen-Ling Kuo and Jane Yung-jen Hsu
Department of Computer Science and
Information Engineering
National Taiwan University
No. 1, Sec. 4, Roosevelt Rd., Taipei 10617, Taiwan

Abstract

Robust intelligent systems may leverage knowledge about the world to cope with a variety of contexts. While automatic knowledge extraction algorithms have been successfully used to build knowledge bases in English, little progress has been made in extracting non-alphabetic languages, e.g. Chinese. This paper identifies the key challenge in instance and pattern extraction for Chinese and presents the *Coupled Chinese Pattern Learner* that utilizes part-of-speech tagging and language-dependent grammar rules for generalized matching in the *Chinese never-ending language learner* framework for large-scale knowledge extraction from online documents. Experiments showed that the proposed system is scalable and achieves a precision of 79.9% in learning categories after a small number of iterations.

Introduction

The continuously evolving content on the Internet can serve as a diverse source of valuable knowledge harvested by domain-independent information extraction softbots (Etzioni et al. 2004). However, training information wrappers requires a significant amount of labeled data, which limits its scalability and diversity.

With the rapid growth of international users and content on the Internet, knowledge extraction of non-English content is becoming increasingly important. However, existing knowledge extraction algorithm do not perform well for non-alphabetic languages, e.g. Chinese, largely due to problems regarding text segmentation, named entity recognition, and different grammars.

This paper presents *Coupled Chinese Pattern Learner* (CCPL) to utilize part-of-speech tagging and finite state machine to solve these problems in the learning process. We incorporate CCPL with the state of art knowledge extraction framework NELL (Mitchell et al. 2015) and started extracting knowledge with an initial ontology defining 145 categories. After 5 iterations, we extracted 4,226 new facts from 114 categories using the ClueWeb09 dataset (Project 2013). The experiment results are evaluated against human-labeled ground truth and showed a precision of 79.8%.

The contributions of this work are summarize as follows. (1) Our work is the first to perform automatic knowledge extraction without domain dependencies in Chinese. (2) It is also the first large-scale automatic Chinese knowledge extraction. Finally, (3) The design of language-dependent instance extractor engine in CCPL can be easily reused in other languages.

Related Work

The pattern based knowledge extraction approaches have shown impressive result in specific domains (Brin 1999; Agichtein and Gravano 2000; Etzioni et al. 2004). Some approaches apply patterns on semi-structure data such as webpages and corresponding tags (Wang and Cohen 2007). Different from one-time extraction, bootstrapped semi-supervised learning leverages the extracted knowledge to improve learned models continuously. However, after many iterations, bootstrapped learning usually suffers from semantic drift, where labeling errors accumulate and make learned concept drift to unrelated concepts (Curran, Murphy, and Scholz 2007).

Never Ending Language Learner (NELL) proposed coupled constraints which applies co-training theory (Blum and Mitchell 1998) to effectively restrain semantic drift. NELL's proposed architecture is also scalable and easy to incorporate different knowledge extraction algorithms. This paper adopts NELL's architecture to extract Chinese knowledge.

Architecture

Chinese NELL (ChNELL) adopts the NELL framework to incorporate language-dependent subcomponents: Coupled Chinese Pattern Learner (CCPL) and Coupled Set Expander for Any Language (CSEAL). Figure 1 shows the architecture of ChNELL.

CCPL is a pattern based knowledge extractor revised from Coupled Pattern Learner. CSEAL is a wrapper based extractor with coupled constraints. CCPL and CSEAL can extract instances independently. The extracted instances are candidates of the knowledge base. When a candidate has high confidence score or is promoted by CCPL and CSEAL at the same time, it can be promoted to be a fact of the knowledge base.

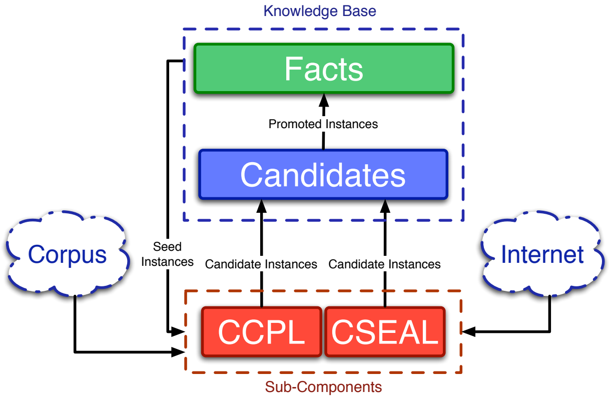

Figure 1: System architecture of ChNELL.

Coupled Constraints

Coupled training is a semi-supervised learning method which couples multiple functions to constrain the learning problem. These functions are "coupled constraints". To ensure the quality of learned results, the learner filters out candidates which violate coupled constraints and have multiple learners vote to generate the most confident candidates. This has been proved to be an efficient method to reduce semantic drift.

Coupled Chinese Pattern Learner

Since CCPL is a revised version of CPL for non-alphabetic languages, we introduce the language independent **Concept Selection** part here and leave the enhanced language dependent **Concept Extraction** part in the next two sections.

Concept selection happens after we extract a great amount of candidate instances and patterns. It involves filtering, ranking, and promotion steps.

Filtering We use mutual exclusion to filter out candidates extracted by two different categories. For example, if "台北(Taipei)" is a candidate instance from predicates "City" and "Province", the instance will be rejected because City and Province are mutual exclusive.

Ranking The ranking mechanism of CCPL is different from the CPL. Instead of ranking candidates by the estimated precision, CCPL ignores candidates which occur in only one promoter and ranks candidates by multi-field sorting.

The first field is diversity. For a candidate i, CCPL evaluates its diversity as follows:

$$diversity(i) = |\{p : p \in P_i\}|$$

, where P_i is a set of patterns which co-occur with instance i.

The second field is the sum of number of patterns co-occured with instance i. It is evaluated as follows:

$$frequency(i) = \sum_{p \in P_i} count(i, p)$$

The function $count(i, p)$ is the times the pattern p co-occurs with instance i in corpus.

We consider that diversity of promoters is more important than frequency of promoters because the higher co-occurrences may indicate a conventional natural language usage or an unfiltered template literal string.

Promotion CCPL promotes top m category candidates. In this paper, $m = 30$. The confidence score of instance is $1 - 0.5^c$, where c is the number of promoters co-occurred with this candidate.

Coupled Set Expansion for Any Language

CSEAL is a component coupled with CCPL. It calls SEAL as its subroutine. SEAL accepts ontology and semi-structured documents as input and extracts instances with similar language structures in paragraphs. SEAL is a language-independent component because it not only looks at text but also meta-data in documents.

Problem of Knowledge Extraction in Chinese

As discussed in previous section, candidate extraction is a language dependent operation in CPL. For knowledge extraction in English, CPL predefines part-of-speech (POS) heuristics to limit instance candidates to be noun phrases and pattern candidates to be meaningful to the target category. Unfortunately, the POS heuristics do not transfer well to all other languages. The challenge in instance extraction and pattern extraction of non-alphabetic languages are explained below.

Instance Extraction

Unlike English, non-alphabetic languages such as Chinese and Japanese do not have natural word delimiters, e.g. space. Word segmentation becomes a necessary step for these languages and it inevitably decreases the accuracy/precision of most language-dependent algorithms.

The first challenge is to recognize unknown noun phrases, which will be segmented into smaller legitimate units when segmenter does not recognize them. Correct identification of noun phrases becomes difficult because numerous possible parsing trees exist when segmented tokens increase. The problem is especially hard for Chinese as new noun phrases are often created by combining known noun phrases.

The second challenge is incorrect word segmentation due to ambiguity. For example, 台北市長庚醫院(Taipei Chang Gung Medical Foundation) is a proper noun which may be incorrectly segmented into "台北(Taipei) / 市長(Mayor) / 庚醫院(Gung Hospital)." Therefore, it is difficult to capture named entities using the segmenter and POS tagger alone.

Pattern Extraction

In general, pattern extraction needs to capture meaningful and informative fragments of sentences in languages. However, the predefined rules in NELL only match very few Chinese sentences because a Chinese sentence may remain valid and syntactically correct after many words are omitted. For example, the words in any pairs of parentheses can be omitted in the following sentence while maintaining its rough meaning: “(我)看到((一)個)(又)高大(又)英挺(的)籃球員(迅速(的))上籃。” (I saw a tall and handsome basketball player quickly layup). So, it is not practically feasible to enumerate all possible POS tag sequences to generate all the matching patterns.

Methodology

Instead of using predefined POS tag sequences, we define the two desired properties a good extracted patterns should follow.

- The captured pattern must be related to the instances of a predicate. Then, we can find more instances in the same predicate by the captured pattern.
- The length of a useful pattern should be within a proper range. For avoiding learned pattern which too specific for some instances or too general.

Verbs and Nouns What type of pattern is informative and meaningful? We observe that instances belonging to the same predicate share a set of verbs. For example, “斷交(severed diplomatic relations)” in Chinese is a single verb which is specific for describing relation between countries. When “斷交” occurs in a sentence, the nearby noun phrases must be any of the country names. Therefore, a good pattern should contain at least one verb which is meaningful to the instance categories.

Finite State Machine of Grammar How many tokens should be extracted to form a valid syntax pattern or an instance? Since it is not feasible to enumerate all possible POS tag sequences using predefine rules. We extract valid pattern (or instance) by finite state machines (FSM) which capture grammatical POS tag sequences. And in our paper we represent finite state machine by regular expression, the POS tags we use are by Stanford POS Tagger.

- **Category Instance:** When the promoted category pattern matched, CCPL identify a noun phrase from the location of wild-card of matched pattern.

  The noun phrase is sequences of adjective with a complementizer and sequences of noun e.g., “英勇強壯的美國隊長(Strong heroic Captain America)”,

  According to the above rules, we summarize as following:

  $$((ADJ(C)?)*|(AP))?((N)+(PU)?)+$$

- **Category Pattern:** When the promoted category instance has been found, CCPL looks forward for candidate patterns from matching instance if there exist arbitrary adverbs followed by at least one verb and a optional preposition e.g., “曾就讀於X(studied at X)”.

  CCPL also allows an optional noun phrase at the forefront e.g., “歌曲收錄於X(songs included in$X$)” or an optional noun phrase followed by a punctuation e.g., “學校，位於X(university, located at X)”.

Predicate Name	#	%
中國皇帝(Chinese emperor)	71	91.5
中藥(Chinese medicine)	72	95.8
公園(Park)	76	47.3
國家(Country)	159	95.6
天氣現象(Weather)	55	90.9
寺廟(Temple)	81	75.3
導演(Director)	52	67.3
島嶼(Island)	68	41.1
情緒(Emotion)	94	95.7
捷運站(MRT station)	111	96.4
政府機構(Government organization)	93	80.6
政治職位(Political position)	71	87.3
方位(Direction)	67	67.1
服裝(Clothing)	87	89.6
歌手(Singer)	123	100
單位(Unit)	106	77.3
疾病(Disease)	74	93.2
程式語言(Programming language)	61	85.2
節慶(Festival)	67	73.1
罪行(Crime)	117	86.3
花草(Flower)	59	96.6
行政區(Administrative division)	131	96.9
街道(Street)	88	93.1
詩人(Bard)	77	98.7
調味品(Condiment)	62	88.7
貨幣(Currency)	105	49.5
資訊工程領域(CS field)	74	55.4
身體部位(Body part)	54	92.5
醫院(Hospital)	96	37.5
銀行(Bank)	68	75
顏色(Color)	95	95.7
鳥(Bird)	54	66.6

Table 1: The number of promoted instances and the corresponding precision on the categories which are extracted over 50 instances after 5 iterations.

  CCPL looks backward for candidate patterns if there are arbitrary adverbs, verbs, preposition and an optional noun phrase e.g., “X進軍好萊塢(X advance to Hollywood)”.

  The finite state machines are show as following:

  $$((N)*|(N)+PU)(AD)*(V)+(PP)?$$

  for looking forward and

  $$(AD)*(V)+(PP)?(N)*$$

  for looking backward.

Experimental Evaluation

The ClueWeb 09 dataset contains 177,489,357 Chinese web pages. The pre-process is as follows. First, HTML tags, JavaScript code blocks and CSS blocks are removed. Second, the sentences are segmented and tagged with part-of-speech by the Stanford natural language processing tool [1]. Third, the short sentences, the sentences without verbs,

[1] http://www-nlp.stanford.edu/

	# of Extracted Instances(%)	# of Correct Instance(%)	Precision
CSEAL	672 (15.9%)	642 (19%)	95.5%
CCPL	3,471 (82.1%)	2,656 (78.6%)	76.5%
Both	83 (1.9%)	77 (2.2%)	92.7%

Table 2: The comparison of the extracted instances from CCPL and CSEAL.

and the sentences with too many punctuation marks are filtered. Finally, the duplicate sentences are deleted. The inputs of CSEAL are documents which are collected by querying search engine on-line. The top 50 related web pages for each permutation are treated as inputs.

The initial ontology consists of 145 categories. The categories includes different types of locations and different types of celebrities.

Results

After 5 iterations, the ChNELL framework generates 4,226 unique instances. The instances are verified by human labeling with agreement, the number of correct instances is 3375, precision is 79.8%. Table 1 list the number of promoted instances and the corresponding precision. Since the number of categories is large, we only show the categories which are extracted over 50 instances after 5 iterations.

To understand the performance of CCPL and CSEAL respectively, we compare the results of them as shown in Table 2.

Discussion

In category extraction, 114/145 categories obtain new knowledge, these categories which with over 60% precision are 96/114, even we raise the precision to 80%, they are still 67/114 categories satisfied this requirement.

In these categories which had good performance, they satisfied at least one of two properties. The first is mutual exclusion did work, categories such as “國家(Country)” with “州或省(StateOrProvince)” and “縣市(CountyOrCity)”, or “歌手(Singer)” with “演員(Actor)” and “導演(Director)”.

The second is the category is a clear concept and no blurring space, such as “詩人(Bard)”, “中藥(ChineseMedicine)”, “顏色(Color)”, “街道(Street)”, “中國皇帝(ChineseMonarch)”, “疾病(Disease)” and “程式語言(ProgrammingLanguage)”. The programming language is interesting one because we did not expect to learn any instance which is not composed of Chinese words. There is no translation for programming language, this proves that CCPL has the ability to adapt to another language in addition to Chinese.

We also have some discoveries for these poor performance categories. The major mistakes can be divided into three types:

- **Fuzzy Concept:** The category is not well defined or it is not only one category which has such functionality. For example, in category “公園(Park)”, we learned instances such as “花市(Flower Market)” and “華山文化園區(Huashan Creative Park)”, the second one actually is an space for exhibition, but they might have the similar functionalities compared with park.
- **Failure of Mutual Exclusion:** Some categories drift to related categories which are close in first impression. For example, Category “海灣(Bay)” drifts to variety of types of boat such as “船隻(Ships)”, “中國輪船(Chinese Steamship)” and “貨輪(Freighter)”.
- **POS Tagging Errors:** The third type of poor performance is caused by instance captured grammar rule. We simply concatenate the a series of nouns to form a instance, when tagger goes wrong, instances also go wrong. For category “島嶼(Island)”, we obtain “蘭嶼東清灣(Orchid Island Tung Ching Bay)” and “綠島柚子湖(Green Island Grapefruit Lake)”.

We can fix the first two problems by choosing predicates and seed instances carefully to maximize the effect of mutual exclusion and easier propagation. For errors due to wrong POS tag, it might be fixed by studying more sophisticated strategy or development of unsupervised learning for general domain named entity recognition algorithm in Chinese.

Conclusion

This paper proposes the ChNELL framework for automatic knowledge extraction of general concepts in Chinese. Like NELL, the ChNELL framework iteratively learns new knowledge by bootstrapping with coupled semi-supervised learning of online documents. The proposed framework extends NELL to handle non-alphabetic languages, e.g. Chinese, by pre-processing the corpus with POS tagging, and formulating language-dependent grammar rules as finite-state machines to capture valid POS tag sequences. Evaluation with 177 million Chinese pages from ClueWeb09 corpus showed that ChNELL is scalable and outperforms existing systems in extracting categories and relations in Chinese.

References

Agichtein, E., and Gravano, L. 2000. Snowball: Extracting relations from large plain-text collections. In *In Proceedings of the 5th ACM International Conference on Digital Libraries*. JCDL.

Blum, A., and Mitchell, T. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory (COLT 1998)*. ACM.

Brin, S. 1999. Extracting patterns and relations from the world wide web. Technical Report 1999-65, Stanford InfoLab. Previous number = SIDL-WP-1999-0119.

Curran, J. R.; Murphy, T.; and Scholz, B. 2007. Minimising semantic drift with mutual exclusion bootstrapping. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACL 2007)*.

Etzioni, O.; Cafarella, M.; Downey, D.; Popescu, A.-M.; Shaked, T.; Soderland, S.; Weld, D. S.; and Yates, A. 2004. Methods for domain-independent information extraction from the web: An experimental comparison. In *Proceedings of the 19th national conference on Artifical intelligence (AAAI 2004)*. AAAI.

Mitchell, T.; Cohen, W.; Hruschka, E.; Talukdar, P.; Betteridge, J.; Carlson, A.; Dalvi, B.; Gardner, M.; Kisiel, B.; Krishnamurthy, J.; Lao, N.; Mazaitis, K.; Mohamed, T.; Nakashole, N.; Platanios, E.; Ritter, A.; Samadi, M.; Settles, B.; Wang, R.; Wijaya, D.; Gupta, A.; Chen, X.; Saparov, A.; Greaves, M.; and Welling, J. 2015. Never-ending learning. In *Proceedings of the Conference on Artificial Intelligence (AAAI) 2015*. AAAI.

Project, T. L. 2013. The clueweb09 dataset.

Wang, R. C., and Cohen, W. W. 2007. Language-independent set expansion of named entities using the web. In *Proceedings of IEEE International Conference on Data Mining (ICDM 2007)*. ICDM.