The 26th International Conference on Legal Knowledge and Information Systems
BOLOGNA - ITALY, 11-13 DECEMBER 2013

UNSUPERVISED KEYWORD EXTRACTION FOR JAPANESE LEGAL DOCUMENTS

Tho Thi Ngoc Le, Minh Le Nguyen, Akira Shimazu
School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
{tho.le, nguyenml, shimazu}@jaist.ac.jp

Bologna, December 13th, 2013

CONTENTS

- Introduction
- Related Work
- Proposed Approach
  - Japanese Linguistic Knowledge
  - Weights
  - Thresholds
  - Post-processing
  - Extracting Keywords
- Experiments
- Conclusions

INTRODUCTION

- Legal documents’ characteristics:
  - Sentences are long and complex;
  - Vocabularies are unique to legal field;
  - Having references within a document or among a set of documents;
  - Document style is also important.

Reading legal documents is **not easy**, especially for non-specialists.

- This study addresses the problem of **extracting keywords** from Japanese legal documents.

INTRODUCTION

- Keywords: words that provide the main ideas or the important content of a sentence or a document.
- Automatic keyword extraction contributes to many applications.
- In Natural Language Processing:
  - Automatic summarization;
  - Text classification;
  - IR, QA systems.
- In legal engineering:
  - Creating indexes for legal cases;
  - Supporting the prediction roles based on legal concepts.

Problem:
Use *unsupervised* approach to extract Japanese legal keywords for the purpose of:

- Summarizing
- Index creating

RELATED WORK

- Keyword extraction supports many applications in legal informatics:
  - Indexing legal cases:
    - (Brüninghaus & Ashley, 1999)
  - Legal concept extraction for predictive roles:
    - (Moens & Angheluta, 2003)
    - (Ashley & Brüninghaus, 2003)
    - (Grabmair & Ashley, 2011)

RELATED WORK

- Two main approaches: supervised and unsupervised
- **Supervised keyword extraction**: treats extraction as the problem of classification to decide if a word is a keyword or not.
- There are a variety of approaches to train the classifiers:
  - (Turney, 1999, 2000): combined heuristic rules with a generic algorithm;
  - (Frank, 1999): used Naïve Bayes method;
  - (Hulth, 2003): added syntactic features (n-grams, part-of-speech tags, NP chunks) to improve performance.

☹ Disadvantage: Need annotated data for training process

- Costly;
- Time consuming.

RELATED WORK

- **Unsupervised keyword extraction**: contains two main steps
  - Collect as many candidate words as possible;
  - Apply a pattern of *(adj + nouns)* to combine candidates and obtain keywords.
- Some remarkable studies:
  - Graph-based ranking approach: find high ranked vertices on text graph based on an assumption that a word is important if:
    - It *connects to as many other words* as possible; or
    - It *connects to important words*.
  - TextRank (Mihalcea & Tarau, 2004) and its variations:
    - SingleRank, ExpandRank (Wan & Xiao, 2008);
    - Degree-based Ranking (Litvak & Last, 2008);
    - Topical PageRank (Liu et. al., 2010).
  - Cluster-based approach: finds the exemplars of clusters to serve as candidates (Liu et. al., 2009).

RELATED WORK

- There is a limited number of studies in extracting Japanese keywords by supervised approaches:
  - (Suzuki et. al., 1997, 1998): combined term weighting and domain identification to train the classifiers;
  - (Ogawa & Matsuda, 1997): segmented Japanese text and extracting the overlapping segments as keywords;
  - (Mathieu, 1999): re-implemented the supervised approach of (Turney, 1995) for Japanese text.
- The disadvantages in Japanese keyword extraction:
  - Applied supervised approaches → Need annotated data for training process;
  - Limited performance;

APPROACH – Japanese Linguistic knowledge

- Japanese writing system consists of four alphabets:
  - Hiragana (ひらがな);
  - Katakana (カタカナ);
  - Kanji (漢字);
  - Romaji (Roman characters).
- Japanese words are grouped in **chunks** (bunsetsu, 文節):
  - A chunk is a block of tokens, so called phrasal unit or bunsetsu segment.
  - A chunk is the smallest inseparable group of tokens in a sentence.
  - A chunk usually includes:
    - One ***independent word***: have meaning and can standalone in a sentence;
    - (Optionally) zero or more than one ***auxiliary word*** to serve the grammatical roles.

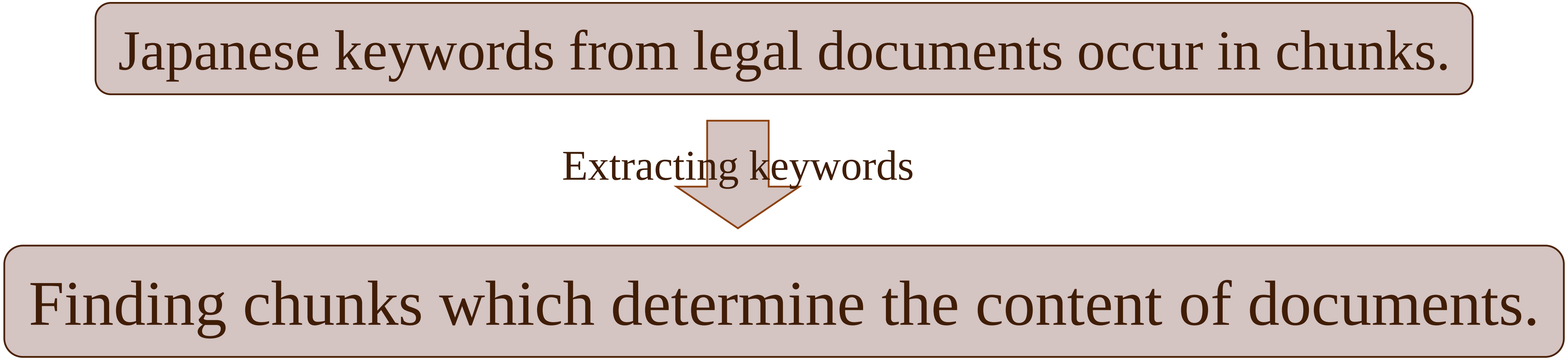
APPROACH – Japanese Linguistic knowledge

- Example for Japanese chunks

この / 法律に / 規定する / 社会保険庁長官の / 権限の / 一部は、 / 政令の / 定める / ところにより、 / 地方社会保険事務局長に / 委任する / ことが / できる。

(Part of the authority of the Chief of a Local Social Insurance Bureau as provided by this Act, and pursuant to the provisions of a Cabinet Order is that they are able to be delegated to the Director General of the Social Insurance Agency.)

- By observation, we found that:

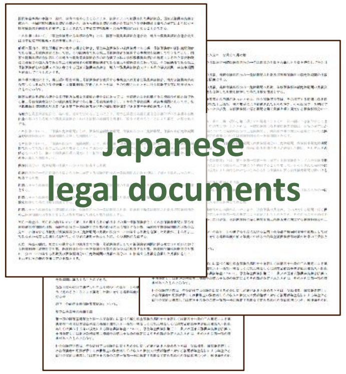


Japanese keywords from legal documents occur in chunks.

Extracting keywords

Finding chunks which determine the content of documents.

APPROACH – Overview



Parse all sentences for **chunks** and **tokens**

この / 法律に / 規定する / 社会保険庁長官の / 権限の / 一部は、 / 政令の / 定める / ところにより、/ 地方社会保険事務局長に / 委任する / ことが / できる。

Assign TFIDF to tokens and **compute chunks' weight**

地方	社会	保険	事務	局長	に	Average
0.8	0.81	2.38	0.77	0.75	0.14	**0.94**

Collect high weighted chunks as candidate for keywords

Remove auxiliary words from candidate chunks

Find the **high weighted chunks** in legal documents.

Keywords

規定
社会保険庁長官
政令
….

APPROACH – Weights of Tokens

- Weight of a token t is TF-IDF score, in which:
  - TF is denoted for Term Frequency which expresses the importance of token within a document d.
    - Raw TF f_t: The number of occurrences of the token in a given documents d;
    - Log TF $tf_{t,d}$: Computed as the logarithmic scaled frequency

$$tf_{t,d} = \log(f_t + 1)$$

  - IDF denoted for Inverse Document Frequency which indicates the importance of token in a collection of N documents D:

$$idf_{t,D} = \log \frac{N}{df_t}$$

  in which, document frequency df_t is the number of documents that contain t.

  - TF-IDF score of a token is the product of its TF and IDF scores:

$$tfidf_{t,d} = tf_{t,d} \times idf_{t,D}$$

APPROACH – Weights of Chunks

- Weight of c is calculated as the average of all tokens included in c:

$$weight_c = \frac{\sum_{t \in T} weight_t}{|T|}$$

in which:

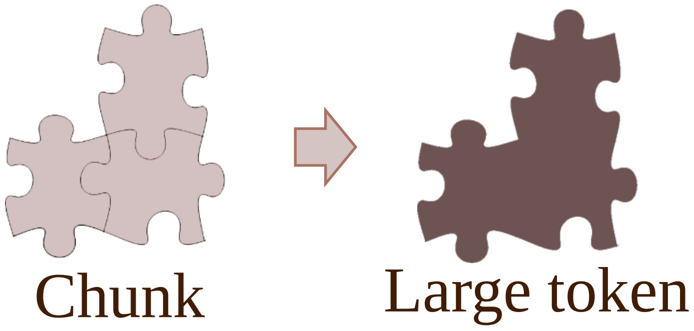
  - $weight_c$ is weight of chunk c;
  - $weight_t$ is weight of token t ($t \in T$);
  - T is the set of all tokens t that belong to the chunk c.

- Example: compute the weight of the chunk **地方社会保険事務局長に**

Token	地方	社会	保険	事務	局長	に	**Average**
Weight	0.80	0.81	2.38	0.77	0.75	0.14	**0.94**

APPROACH – Thresholds

- Important tokens have higher weights compared to the average weight of the other tokens in a document.
- If we view:



candidates
=
high weight tokens

➡ A chunk is candidate for keyword if it has high weight.

- Keywords can contain one or more tokens:
  - Single keywords: contain a single token;
  - Compound keywords: contain multiple tokens.
- Two thresholds are designed for two kinds of keywords.

Token
Chunk
Token
Token
Token

APPROACH – Thresholds

- Threshold θ_c for compound keywords

$$\theta_c = \beta \times \frac{\sum_{t \epsilon T_d} weight_t}{|T_d|}$$

where:

- $weight_t$ is weight of token t;
- T_d is the set of all tokens appearing in the given document d;
- $|T_d|$ is the number of tokens in T_d;
- β is coefficient to control the number of compound keywords.

APPROACH – Thresholds

- Threshold θ_s for single keywords

$$\theta_s = \alpha \times \frac{\sum_{t \epsilon T_d^+} weight_t}{|T_d^+|}$$

where:

- $weight_t$ is weight of token t;
- T_d^+ is the set of all tokens whose *weights are greater than zero* appearing in the given document d;
- $|T_d^+|$ is the number of tokens in T_d^+;
- α is coefficient to control the number of single keywords.

APPROACH – Post-processing

- Keywords must have **meaning**;
- Keywords must be **independent words;**
- Keywords must **not contain auxiliary words** severing for grammatical roles;

➡ Remove unnecessary words from the beginning and ending of candidate chunks;

APPROACH – Extracting Keywords

- Collect candidate chunks whose weights are higher than threshold θ_c;
- Post-process the candidates for compound keywords;
- The keywords which contain only one token should have weights higher than θ_s;

EXPERIMENTS

- Data: The Japanese National Pension Act (JNPA) includes:
  - 879 sentences;
  - 208 keywords.
- Tool for decomposing Japanese sentences: Cabocha[1]
- Corpora for calculate IDF:
  1. 7,984 legal documents obtained from Japanese government webpage.
  2. 496,997 news articles from Mainichi Shimbun newspaper (1991-1995).
- Combinations of corpora to calculate IDF:
  3. All 7,984 legal documents and all 496,997 news articles;
  4. All 7,984 legal documents and all 11,497 news articles in year 1995;
  5. All 7,984 legal documents and 7,984 news articles in year 1995;

[1] Taku Kudo and Yuji Matsumoto, “Japanese Dependency Analysis using Cascaded Chunking,” in Proc. Conf. ACL-CoNLL ’02, Taiwan, 2002, pp. 63-69.

EXPERIMENTS

- Calculate TFIDF with:
  - Raw TF; and
  - Logarithmic scaled TF.
- $\beta = 1; \alpha = 1.$
- POS tags of independent words:

ChaSen POS Tag	Description	Example	ChaSen POS Tag	Description	Example
名詞-一般	Noun(general)	耳(ear)	名詞-形容動詞語幹	Noun(adjective -na)	安全(safe)
名詞-固有名詞-一般	Noun(proper.general)	光が丘(Hikarigaoka)	名詞-接尾-一般	Noun(suffix.general)	印(mark)
名詞-固有名詞-人名-一般	Noun(proper.name.general)	お市の方(Oichinokata)	名詞-接尾-地域	Noun(suffix.place)	駅(station)
名詞-固有名詞-人名-姓	Noun(proper.name.surname)	山田(Yamada)	名詞-接尾-サ変接続	Noun(suffix.verbal)	話(story)
名詞-固有名詞-人名-名	Noun(proper.name.firstname)	紀子(Noriko)	名詞-接尾-形容動詞語幹	Noun(suffix.adjective-na)	-的(-tive)
名詞-固有名詞-組織	Noun(proper.organization)	NHK	形容詞-自立	adjective–i(free)	苦い(near)
名詞-固有名詞-地域-一般	Noun(proper.place.general)	京都(Kyoto)	形容詞-非自立	adjective -i(bound)	難い(difficult)
名詞-固有名詞-地域-国	Noun(proper.place.country)	日本(Japan)	形容詞-接尾	adjective -i(suffix)	っぽい(like)
名詞-非自立-一般	Noun(bound.general)	こと(thing)	接頭詞-名詞接続	prefix(+noun)	両(both)
名詞-サ変接続	Noun(verbal)	見学する(visit)			

EVALUATION

- Evaluation criteria: Precision, Recall, F1-score:

$$P = \frac{\#\ correct\ keywords}{\#\ extracted\ keywords};\ R = \frac{\#\ correct\ keywords}{\#\ annotated\ keywords};\ \mathrm{F1}\_score = 2 \times \frac{P \times R}{P + R}$$

- Compare to the most popular graph-based ranking approach TextRank[2]:
  - Constructing a graph of words from text:
    - Vertices: Words from text;
    - Edges: two words have relations if they occur in a co-occurrence window.
  - Ranking to find the high weighted vertices in the graph;
  - Collecting the high weighted vertices as candidates for keywords;
  - Combining the candidates to obtain keywords.

[2] Rada Mihalcea and Paul Tarau, “TextRank: Bringing Order into Texts,” in Proc. Conf. EMNLP-ACL ’04, Spain, 2004, pp. 404-411.

EVALUATION - Comparisons

Method		# Extr. Keys	# Corr. Keys	Precision	Recall	F1-score
TextRank						
Co-occ.window = 2		918	177	19.28	85.10	31.44
Co-occ.window = 6		693	165	23.81	79.81	**36.68**
Chunk-based approach						
Corpus for IDF	TF					
Mai: ‘91 to ‘95	raw	803	190	23.66	91.35	37.59
Mai: ‘91 to ‘95	log	456	145	31.80	69.71	43.67
Law	raw	532	161	30.26	77.40	43.51
Law	log	266	106	**39.85**	50.96	44.73
Combine all	raw	827	193	23.34	**92.79**	37.29
Combine all	log	463	148	31.97	71.15	44.11
Law & Mai’95 (all)	raw	768	192	25.00	92.31	39.34
Law & Mai’95 (all)	log	458	150	32.75	72.12	45.05
Law & Mai’95 (part)	raw	817	193	23.62	**92.79**	37.66
Law & Mai’95 (part)	log	356	133	37.36	63.94	**47.16**

EVALUATION - Analysis

- Keyword examples on the cases that TextRank gets wrong:

TextRank	Chunk-based	Annotated
Cannot be extracted caused by numerical noun “一” and noun suffix “時” (lump-sum)	死亡一時金	死亡一時金
Cannot be extracted caused by noun suffix “等” (such as)	公的年金被保険者等総数	公的年金被保険者等総数
Cannot be extracted caused by noun suffix “者” (person)	被保険者期間	被保険者期間
ない社団等	No	No

- In TextRank, if we accept noun suffix such as “等” (*means: etc., such as*) in pre-processing step for candidates, then we will extract keywords like: “総額等” and “ない社団等”.

CONCLUSIONS

- Proposed a novel unsupervised approach to extract keywords from Japanese legal documents.
- Characteristics of chunk-based approach:
  - Exploring common knowledge of Japanese chunks;
  - Using basic concept in NLP: TF-IDF;
  - Simple technique to collect candidate words: the average.

However, our method is still effective (improve 10.5% on F1-score).

- In the future, we may explore other methods for weighting Japanese chunks.

THANK YOU !

SUB-SLIDES

- Keyword examples:

Keywords	Meaning
日本私立学校振興・共済事業団	Private schools of Japan
国家公務員共済組合連合会	Federation of Mutual Aid Associations of National Public Service Personnel
地方公務員共済組合連合会	Pension Fund Association for Local Government Officials
厚生労働省令	The Ordinance of the Ministry of Health, Labour and Welfare
国民年金事業	National Pension service
…	…

SUB-SLIDES

- ***Independent words*** (analogous to *free morphemes* in English) are words which have meaning and can stand alone in a sentence. In Japanese, independent words can be: nouns, pronouns, verbs, adjectives, adjectival nouns, adverbs, adnominal adjectives, conjunctions, or interjections.
- ***Auxiliary words*** are analogous to *bound morphemes* in English. They usually follow independent words to express the variation in meaning or to make clear the relations between and among independent words. In Japanese, auxiliary words can be either auxiliary verbs or particles.