Automatic classification of technical papers by using title analysis

Imai Shun

School of Information Science,
Japan Advanced Institute of Science and Technology

February 15, 1999

Keywords: classification of technical papers, title analysis, functional phrase, terminological dictionary.

As a large number of technical papers is available in the form of digital library and electronic newspaper these days, it becomes important to provide easy access to the papers. The information retrieval systems using keyword are widely available, however these systems do not always lead satisfactory results. On the other hand, most people can easily find papers of interest if they are already classified, however, it is difficult and time-consuming to classify a great deal of papers by hand. Therefore, automatic classification becomes necessary.

Many of the researches on automatic classification of electronic texts including academic papers have used statistically based techniques. These techniques have been proven to be effective and are extensible. In the mean time, people who categorize the papers often prefer summary information such as titles. Especially, an expert only need to take a short look at a paper's title of his field to estimates its purposes and determines the proper classification category. In case of the expert of artificial intelligence, as he reads "プロダクション・システムによる線画の解釈 (Line drawing interpretation by using a production system)", he estimates that its subject is "線画の解釈 (Line drawing interpretation)" and its method is "プロダクションシステム (production system)", because of two reasons:

1. The expert has enough knowledge about domain-specific terms on his field.

2. The title is the digest of a paper and contains technical terms.

I assume that these two reasons are true, so technical papers can be correctly classified only by using title analysis, if a rich terminological dictionary is available.

This paper proposes the automatic classification method of Japanese technical papers by using a terminological dictionary and title analysis.

The method consists of **Structural Mapper** and **Term Detector**. Structural Mapper divides a technical paper's title into a few components by applying mapping rules repeatedly. Mapping rules are as follows.

1. Omit unimportant parts such as system names, tails, proper names, etc.
   (e.g.)MBT1:実例に基づく訳語選択 (MBT1:Example-based word-selection)

2. Divide by string pattern matching.
   (e.g.) XとそのY (X and its Y) → X, そのY (its Y)[1]
   (e.g.) 集合束縛変数とその自然言語処理への応用 (Set Bound Variables and Its Application to Natural Language Processing) → 集合束縛変数 (Set Bound Variables), その自然言語処理への応用 (Its Application to Natural Language Processing)

3. Transform by string pattern matching to the standardized structure.
   (e.g.) XのYへの応用 (An application of X to Y) → Xを応用したY (Y by using X)
   (e.g.) 知識工学の機械設計CADへの応用 (The Application of Knowledge Engineering to Machine Design CAD) → 知識工学を応用した機械設計 CAD(Machine Design CAD by using The Application of Knowledge Engineering)

4. Omit unimportant words such as の研究 (study), の方法 (method) .
   (e.g.) モデル事例ベースを用いた定性的多目的最適設計に関する研究(A Study on Qualitative Multi-Objective Optimum Design Using Model Case Base)

5. Divide by using functional phrases such as *wo-mochiita*, *ni-yoru*.
   (e.g.) X 機能語 Y (X functional phrases Y) → X, 機能語 (functional phrases), Y
   (e.g.) 知識を用いた建築図面の理解 (Understanding Architectural Drawings Using Knowledge) → 建築図面の理解 (Understanding Architectural Drawings), use, 知識 (Knowledge)

Term Detector extracts technical terms that are contained in compound nouns by using a terminological dictionary, which consists of technical terms and their codes. This system uses the term-tree of computer science in Encyclopedic Dictionary of Computer Science[2] as terminological dictionary. Term Detector uses the following rules when it extracts technical terms from compound noun phrases.

1. Find the longest technical term by backward matching from a compound noun phrase.
   (e.g.) 簡単なパルス回路(simple pulse circuits)

2. If the last word of the compound noun phrase is an omissible such as "system" and "mechanism", omit it, and then apply rule 1.
   (e.g.) 高速仮説推論システム (a fast hypothetical reasoning system)

[1] X, Y are compound nouns
[2] In Japanese. Iwanami Publisher, 1990

3. When the compound noun phrase contains case markers, such as *wo* and *no*, omit them, and apply rule 1.
   (e.g.) エキスパートシステムの構築環境(an developing environment of expert systems)

4. When the compound noun phrase contains "的 (like)" and "型 (type)", skip them, and apply rule 1.
   (e.g.) 概念の学習(learning of concept) ··· 概念学習 (concept learning)

5. When the compound noun phrase contains **サ変名詞** (sahen-noun) or "**名詞** + 化 (noun + -ization)", omit them, and apply rule 1.
   (e.g.) 分散型探索機構 (distributed search mechanism)

Then, the system assigns the matching code for the domain-specific terms that are obtained by above rules.

To evaluate the performance of this method, I used 369 Japanese papers which are in Journal of Japanese Society for Artificial Intelligence. The classification codes of 292 papers was the same as the codes that are assigned by human; the method achieved 79% accuracy. For the remaining 77 papers, the reasoning method is necessary to assign appropriate classification codes. Most of other errors are caused by lack of terms in the dictionary, and the morphological analysis.