翻訳辞書を活用した日中クロスリンガル検索システムの試作と評価

大倉清司、山下達雄、富士秀、徐国偉、長瀬友樹、潮田明

（株）富士通研究所

okura.seiji@jp.fujitsu.com

1. はじめに

翻訳辞書を活用した日中クロスリンガル検索を試作し、専門用語辞書の効果および日本語、中国語のストップワードの効果を調査した。検索として実用的なシステムを目指し、形態素解析によるインデックス作成方法で実験を行った。中日翻訳システムから一部を切り出して、形態素解析システムを開発し、また翻訳辞書からキーワード変換辞書を抽出した。専門用語辞書を追加することにより、検索精度が向上し、ストップワードの効果も検証した。

2. 中国語を対象としたクロスリンガル検索システムの研究

NTCIR[1]で検索システムの研究がなされているが、その中でのクロスリンガルセッション（CLIR）において、中国語を対象とした研究が盛んに行われている。

NTCIR4 の大会では、日中のクロスリンガル検索システムに２つのシステム結果が提出された[2]。１つは英語をピボットとしたクロスリンガル検索システム[3]で、ダイレクトに日本語から中国語への検索システムではない。もう１つは翻訳システムを使ったクロスリンガル検索システムである。

日中クロスリンガル検索に関して漢字概念をマッピングさせる方式での検索システムによる評価がある[4]。

3. 日中クロスリンガル検索システム

中国語は英語のように単語間にスペースがないため、形態素解析するにはツールが必要である。本研究では翻訳辞書を活用した日中クロスリンガル検索システムを開発した（図１）。形態素解析システムは中日翻訳システムの一部を切り出して開発したので、専門用語辞書の登録も容易である。キーワード変換辞書は翻訳辞書から抽出した。クロスリンガル検索システムには、入力キーを目標言語に変換して検索する方法と、対象となる文書を検索側言語に翻訳したものを検索する方法などがあるが、我々は前者で実験した。すなわち、検索対象の中国語文書はあらかじめ形態素解析してインデックスをはっておき、検索時には日本語の検索キーを形態素解析し、その結果をキーワード変換で中国語キーワード群に変換して検索するものである。検索手法はベクタースペースモデルに基づき、類似度順にランキングして表示するものである。例えば「中国語検索」という日本語で検索すると、入力キーを「中国語」「検索」と形態素解析し、それぞれの単語を「中文」「汉语」「搜寻」、「检索」「查找」と翻訳辞書を使ってキーワード展開する。キーワード展開された単語のベクトルと、対象文書のベクトルとを比較して、近いものからランキング出力するというものである。

4. システムの試作

ベクタースペースモデルに基づく検索システムを開発した。入力された日本語キーを形態素解析するのには翻訳エンジン ATLAS[5]を改造して、利用した。キーワード変換辞書については、中日翻訳システムの辞書から主に名詞のエントリーを抜き出して作成した。専門用語の効果を調査するため、約一万語を専門用語辞書として追加した。対象文書の形態素解析をするために、

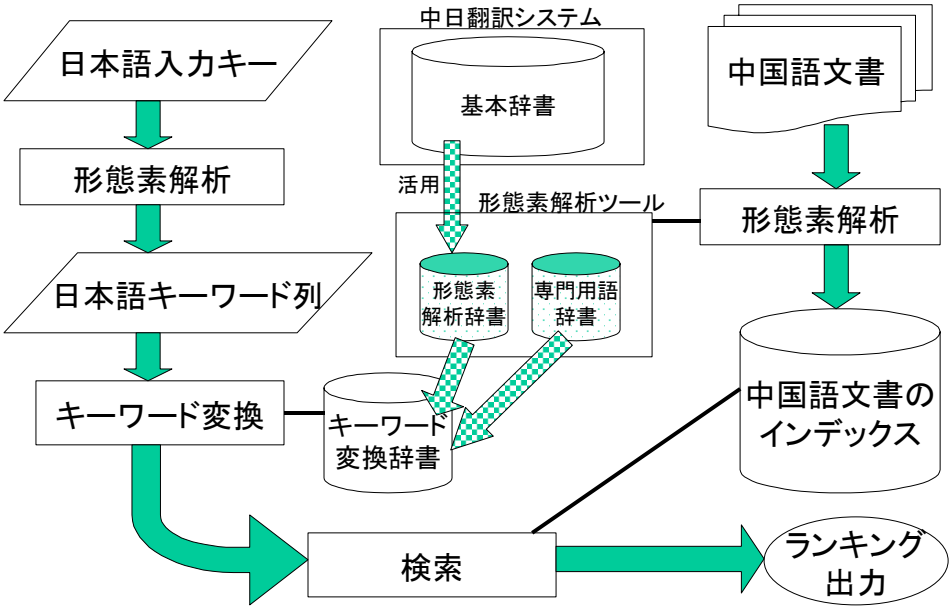

図1：日中クロスリンガル検索システム

中国語の翻訳エンジンから一部を切り出し、形態素解析ツールを開発した。基本となる辞書で形態素解析したものと専門用語辞書も使って形態素解析したものの２通りでインデックスを作成した。形態素解析ツールは専門用語辞書への単語登録も容易にできる。
ストップワードについては、キーワード変換時に日本語のストップワードを変換しないようにするモジュールを作成した。中国語のストップワードに関しては、インデックス作成の際および、キーワード変換の際にストップワードを入れないようにした。

5. 評価実験

開発したシステムで評価実験を行った。「中華人民共和国国務院令」など中国語の法令関連の約 1000 文書を対象に日本語文で検索して、正解が含まれる文書が何位にランキングされるか評価した。
専門用語辞書の効果を試すために、２つの観点から実験を行った。

1. キーワード変換辞書に専門用語を入れたとき、入れないときの比較
2. 中国語文書インデックス作成時に、専門用語辞書を使って形態素解析した場合と基本辞書のみを使って形態素解析した場合の比較

この２つの観点を組み合わせて実験をした。
また、キーワード変換におけるストップワードの影響についても実験した。日本語のストップワードおよび中国語のストップワードを用意し以下の実験を行った。

1. ストップワードを使用しない場合
2. 日本語のストップワードだけ使って検索した場合
3. 中国語のストップワードだけ使って検索した場合
4. 日本語および中国語のストップワードを使って検索した場合

実験をまとめると表１のようになる。

6. 実験結果

実験結果を表２に示す。この結果からわかったことについて考察する。全般的に、

- 基本辞書のみでキーワード変換し、中国語文書を対象にインデックス作成時も基本辞書で解析した場合と、
- 基本辞書＋専門用語でキーワード展開し、中国語文書を対象にインデックス作成時も基本辞書＋専門用語辞書で解析した場合

を比較すると、後者のほうがよい結果が出た。これはどういう要因によるものか、個々の観点について分析する。

キーワード変換の専門用語への影響

専門用語には、例えば以下のようなものが含まれる：

譲渡契約
登録商標譲渡申請
知的財産権保護

キーワード変換辞書の専門用語への影響は、インデックス作成時にも専門用語辞書を使う場合に効果が大きいが、インデックス作成時に基本辞書を使う場合においても、ある程度効果があるようだ。Ａ－１、Ｂ－１の実験結果を比較すると、クエリー５において、専門用語を使わないキーワード展開だと正解文書が 2 位にランキングされたのに対し、専門用語辞書を使ったキーワード展開では正解文書が 1 位にランキングされた。展開された中国語キーワードは専門用語のキーワード変換のほうが１つ多く、その１つが専門用語であった。これにより、検索キーワードのベクトルがより正解文書に近くなったからである。
専門用語辞書を使って対象文書を解析しなくても、キーワード変換辞書に含まれている専門用語とインデックスに含まれる単語が一致すればいい結果が出る。専門用語の用語集があれば、それを使ってキーワード展開したほうがよい結果が得られることがわかる。

インデックス作成時の中国語解析における専門用語の影響

インデックス作成のため中国語文書解析時に、専門用語辞書を使う場合と使わない場合の比較を行った。これは表２のＢ－１、Ｂ－２の結果によるものである。専門用語辞書を使って解析したものでインデックスを作成したほうがいい結果が出た。特に、クエリー６において、基本辞書を使って解析したものは、正解文書が上位５０件にもランキングしなかった。これは、検索の要となる専門用語がインデックスに入っていなかったからである。

ストップワードの影響

・ストップワードの効能に関しては、今回の実験結果から以下のことが言える。

- 日本語のみでストップワード（ひらがなの単語など）を使用したとき

効果なし…ストップワードに入れた単語（ひらがな語など）は変換辞書になく展開されないため。

- 中国語のストップワードのみを使用したとき

効果あり…基本辞書のみでキーワード展開・インデックスも基本辞書のみ、の実験においても、高精度な結果が出た。

- 日本語＋中国語のストップワードを使用したとき

効果は中国語のストップワードのみを使用したときと同じ

検索には tf・idf でベクタースペースモデルに基づいたモデルを使っているが、中国語で頻出する単語（例えば「的」）などはストップワードとしてインデックス作成時にも検索時にもベクトルに含まないようにすると、効果があった。

変換辞書　　インデックス
Ａ:基本辞書　　１:基本辞書で文書を解析
Ｂ:基本辞書＋専門用語辞書　２:基本辞書＋専門用語辞書で文書を解析
※SW＝ストップワード

	SWなし	日本語SWのみ	中国語SWのみ	日中SW
Ａ－１				
Ｂ－１				
Ｂ－２				

表1： 実験の概要

7. まとめ

翻訳辞書を使った日中クロスリンガル検索を開発した。実際的な検索システムを目指し、対象文書を形態素解析するシステムを開発した。今回開発した日中クロスリンガル検索システムで以下のことについて検証した。

- キーワード変換辞書を充実させることにより、対象文書の解析時に専門用語辞書を使わなくてもある程度検索精度の向上が期待できる。
- 今回の検索実験においては、日本語のストップワードは効果がみうけられない。
- 中国語のストップワードは、tf・idfを使った類似度計算においても大変有益である。

今回、正解セットを使ってF-measureなどでの評価はしていない。今回の結果を、正確な精度評価に基づきさらに分析したい。

参考文献

[1] http://research.nii.ac.jp/ntcir/index-ja.html

[2] Kazuaki Kishida, Kuang-hua Chen, Sukhoon Lee, Kazuko Kuriyama et al. *Overview of CLIR Task at the Fourth NTCIR Workshop*. Working Notes of NTCIR-4, 2004.

[3]Tetsuji Nakagawa and Mihoko Kitamura. *NTCIR-4 CLIR Experiments at Oki*. Working Notes of NTCIR-4, 2004.

[4] Md.Maruf Hasan and Yuji Matsumoto. *Japanese-Chinese Cross-Language Information Retrieval: An Interlingua Approach*. Computational Linguistics and Chinese Language Processing Vol.5, No.2, August 2000, pp.59-86..

[5]富士通.英日・日英翻訳ソフト ATLAS. http://software.fujitsu.com/jp/atlas/.

変換辞書　　インデックス
A:基本辞書　　1:基本辞書
B:専門用語辞書2:専門用語辞書

※SW＝ストップワード

凡例

oxxxxxoo	←検索結果1位
ooxooxoo	←検索結果2位
oooooxoo	←検索結果3位
oooooxoo	←検索結果4位
oooooxoo	←検索結果5位
oooooxoo	←検索結果6位

↑左から順に、クエリー1、クエリー2、クエリー3、..クエリー8の結果で、oの場合、その順位までに正解文書がランキングされたことを示す。

	SWなし	日本語SWのみ	中国語SWのみ	日中SW
A－1	oxxxxxoo	oxxxxxoo	ooooooxoo	ooooooxoo
	ooxooxoo	ooxooxoo	ooooooxoo	ooooooxoo
	oooooxoo	oooooxoo	ooooooxoo	ooooooxoo
	oooooxoo	oooooxoo	ooooooxoo	ooooooxoo
	oooooxoo	oooooxoo	ooooooxoo	ooooooxoo
	oooooxoo	oooooxoo	ooooooxoo	ooooooxoo
B－1	oxxxoxoo	oxxxoxoo	ooooooxoo	ooooooxoo
	ooxooxoo	ooxooxoo	ooooooxoo	ooooooxoo
	oooooxoo	oooooxoo	ooooooxoo	ooooooxoo
	oooooxoo	oooooxoo	ooooooxoo	ooooooxoo
	oooooxoo	oooooxoo	ooooooxoo	ooooooxoo
	oooooxoo	oooooxoo	ooooooxoo	ooooooxoo
B－2	oooxxooo	oooxxooo	oooooooo	oooooooo
	oooxoooo	oooxoooo	oooooooo	oooooooo
	oooxoooo	oooxoooo	oooooooo	oooooooo
	oooxoooo	oooxoooo	oooooooo	oooooooo
	oooxoooo	oooxoooo	oooooooo	oooooooo
	oooooooo	oooooooo	oooooooo	oooooooo

表2: 実験結果