制約付きHMMとピボット言語を用いたキーワードリストからの対訳用語抽出

Bilingual Terminology Acquisition from Keyword Lists Using Constrained HMM and a Pivot Language

京都大学大学院情報学研究科 **中澤　敏明**

PROFILE: 2010 年京都大学大学院情報学研究科知能情報学専攻博士課程修了。博士（情報学）。機械翻訳の研究に従事。

nakazawa@nlp.ist.i.kyoto-u.ac.jp　TEL 075-753-5346

京都大学大学院情報学研究科 **Denny Cahyadi**

PROFILE: 2012 年京都大学大学院情報学研究科知能情報学専攻修士課程修了。修士（情報学）。現在はパナソニック株式会社に勤務。

京都大学大学院情報学研究科教授 **黒橋　禎夫**

PROFILE: 1994 年京都大学大学院工学研究科電気工学第二専攻博士課程修了。博士（工学）。2006 年 4 月より京都大学大学院情報学研究科教授。自然言語処理、知識情報処理の研究に従事。

1 はじめに

専門用語の訳語集は、二言語以上を扱う自然言語処理の多くのアプリケーションにおいて重要である。しかし、専門用語は専門性が高く、新語が続々と生成されるため、高精度かつカバレージの高い訳語集を構築するのは難しい。そのため、専門文書から訳語対を自動で獲得する研究に注目が集まっている。

本研究では、先行研究 [1] に倣って、二種類の言語で書かれたキーワードリストを科学技術論文から抽出し、各キーワードの言語間の対応付けをとる。しかし、キーワードリスト対は、キーワードレベルのずれとキーワードを構成する単語レベルのずれを含みうるため、対応付けをとることは容易ではない。我々はこの問題を解決するため、機械翻訳における単語アラインメントモデルとして広く利用されている隠れマルコフモデル (HMM) にいくつかの制約を導入し、教師無しでキーワードを効果的に対応付ける。

一方、利用可能なリソースの規模が小さいなどの理由から、訳語対の獲得が困難な言語対も存在する。この問題に対し、我々は中間言語を用いることで、より多くの言語間において訳語対が獲得できることを示す。本研究では、英語を中間言語として用い、日本語と中国語のキーワードの対応付けを行う。【図 1】に本研究全体の枠組みを例示する。

2 制約付きHMMを用いたキーワードリストからの対訳用語抽出

2.1 制約付き HMM

キーワードリスト対はその論文の著者が書いたものであるため、理想的には同じ個数、同じ順序、完全な対訳として書かれているはずである。しかし実際には、個数や順序が異なったり、訳が不正確であったりするものも少なからず存在する。

先行研究 [1] では、個数が同じキーワードリスト対に対し、先頭から順に対応付けることで対訳用語を抽出している。また個数の異なるキーワードリスト対に対しては、既存の対訳辞書を利用することで、キーワード間の部分対応を発見し、これを手掛かりとして対訳用語を抽出している。キーワードリスト対がクリーンな場合はこの方法による抽出手法で十分であるが、前記のようなノイズの混じったデータに対しては不十分である。また、

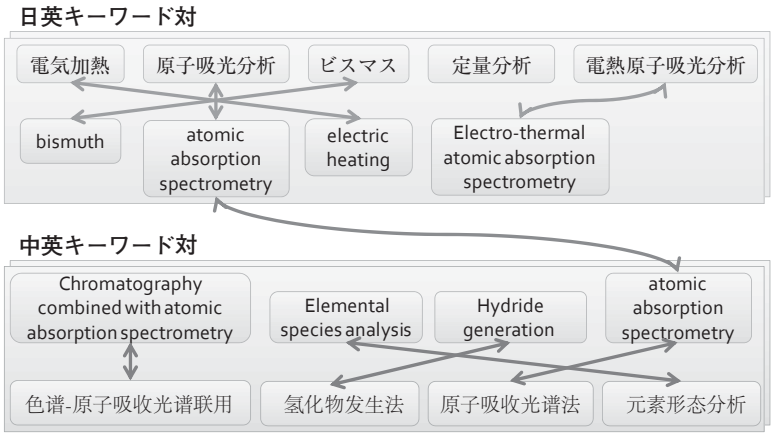

図 1：英語を中間言語とした日英・中英キーワード対からの日中対訳用語抽出

キーワードは多くが専門用語であるため、既存の辞書では部分的な対応ですら発見することは難しい。本研究では、これらの問題を解決するための手法を提案する。

キーワードリスト対からの対訳用語抽出問題は、キーワードを単語、キーワードリストを文とみなせば、機械翻訳などで一般的に行われている対訳文内の単語対応発見（アライメント）問題と同様の方法で解くことができる。しかし各キーワードリストには平均して数個のキーワードしかなく、また各キーワードの出現回数はそれほど多くないため、データスパースネス問題が発生するなど、単語アライメントとまったく同様の方法で高精度な対応付けを行うことは難しい。

本研究では、単語アライメントにおいてよく使われる手法の一つである HMM によるアライメント [2] を拡張し、さらに制約を加えることで、キーワードアライメントに対応できるようにする（モデルの詳細は参考文献 [3] を参照）。このモデルでは、各キーワードに、対応がないことを示す「null」とキーワードの末尾を示す「EOK」という 2 つの特殊な単語を挿入する。また制約として、1. EOK は相手言語の EOK しか生成しない、2. EOK からは EOK 以外のすべての状態に遷移できるが、EOK 以外の単語からは、同一キーワード内の単語にしか遷移できない、という条件を加える。

このモデルは、キーワードレベルとキーワード内の単語レベルの二階層の対応を考慮したものであり、キーワードリストのノイズやデータスパースネスに対して頑健である。【図 2】に「ハイビジョンテレビ , ロボットの視覚」と「high definition TV, robot vision」というキーワード対に対する制約付き HMM によるアライメント例を示す。濃い灰色の部分は遷移確率が 0 の部分であり、薄い灰色の部分は単語出力確率が 0 の部分である。これにより、キーワード内の各単語の対応とキーワードの対応の 2 つが同時に得られる。

2.2 対訳用語抽出実験

制約付き HMM の有効性を示すために、日英および

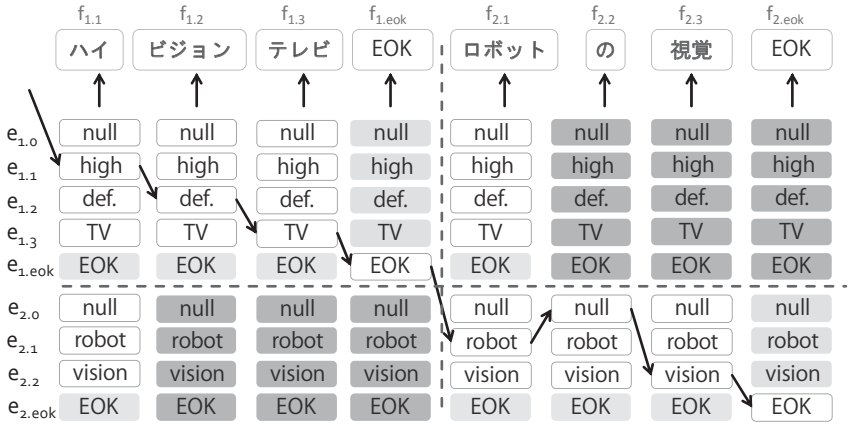

図 2：制約付き HMM を用いたキーワードアライメントの例

中英の対訳用語抽出実験を行った。日英はCiNii（http://ci.nii.ac.jp/）の論文データより抽出した約72万キーワード対をトレーニングに用い、そのうちの100キーワード対に人手で正解の対応を付与し、テストデータとした。キーワードリスト内の平均キーワード数は8.07で、ノイズの含まれるキーワード対の割合は約36%であった。中英はCNKI（http://www.cnki.net/）の論文データより抽出した約7万8千キーワード対をトレーニングに用い、そのうち100キーワード対に正解を付与した。キーワードリスト内の平均キーワード数は4.04で、ノイズの含まれるキーワード対の割合は約6%であった。評価は再現率·適合率·F値により行った。

【表1】に日英の実験結果を、【表2】に中英の実験結果を示す。Monoはキーワードを先頭から順に対応付ける手法による結果、GIZA++[4]は機械翻訳分野で広く使われている単語アライメントツールを用いて、キーワード単位でのアライメントを行った結果、HMMは制約なしの既存のHMMによる結果である。日英では、ノイズの割合が大きいにもかかわらず、全ての指標で最も高い値を示しており、中英においても最も高い適合率を示している。中英のトレーニングデータは日英に比べてトレーニングデータのサイズが1桁小さいため、日英ほどの精度にはなっていないが、データを増やすことでさらに精度を向上することができると考えられる。また中英のデータは日英よりもクリーンなデータであり、Monoの精度がかなり高いが、同様にトレーニングデータを増やすことで、この精度を超えることも可能と考えている。

表1：日英対訳用語抽出実験結果

	再現率	適合率	F値
Mono	89.0	74.2	80.9
GIZA++	92.2	90.6	91.4
HMM	92.8	60.3	73.1
提案手法	95.4	91.8	93.6

表2：中英対訳用語抽出実験結果

	再現率	適合率	F値
Mono	91.8	92.0	91.9
GIZA++	82.0	89.7	85.7
HMM	76.9	41.1	53.6
提案手法	80.6	94.7	87.1

3 中間言語を利用した対訳用語獲得

3.1 対訳誤りの増幅

英語とその他の言語のキーワード対データは比較的容易に手に入るが、たとえば日本語と中国語などの言語対のデータはほとんど存在しないと言ってよい。しかしこのような言語対であっても、英語を中間言語として利用し、日英・中英の対訳用語を結合することで、日中の対訳用語を獲得することが可能である。ここで問題となるのが、結合前の日英・中英それぞれの対訳誤りおよび曖昧性のある対訳による、結合後の日中対訳での誤りの増幅である。

たとえば、前節で獲得した日英・中英対訳のうち結合可能であったものの精度は、日英が約69%（約31,000対）、日中が約81%（約28,000対）であったが、結合後の日中対訳の精度は約32%（約92,000対）にまで低下してしまう。我々はこの問題を解決するために、SVMを利用した結合後の日中対訳のフィルタリングと、結合前のそれぞれの対訳の頻度によるフィルタリングの二つの方法を提案する。

3.2 SVMによる結合後フィルタリング

SVMによるフィルタリングを行う際に利用する素性として、以下に挙げるものを利用した

1. 日英・中英対訳での頻度
2. 英語抄録を利用したキーワードの文脈類似度[5]
3. 日中共通漢字情報によるキーワードの文字類似度[6]

実験として、前記結合後の日中対訳約92,000対のうち800対に人手で正しい対訳かどうかのラベルを付与したデータ（正例437、負例363）を利用し、分類器のトレーニングおよび精度評価に利用した。精度はleave-one-out交差検定により算出した。結果のコンフュージョンマトリックスを【表3】に示す。この表から、再現率78.5%、適合率76.2%、F値77.3%となった。また、ラベルなしデータすべてを分類器にかけたところ、正と判断された対訳対は約22,000対であった。ここからランダムに200対をサンプルして人手で正誤判定したところ、精度は73.5%であり、フィルタリング前の精度32%から大幅に向上することができた。

表 3：SVM によるフィルタリング精度

	正解＋	正解－	合計
SVM+	343	107	450
SVM-	94	256	350
合計	437	363	800

3.3　結合前対訳の頻度によるフィルタリング

結合後対訳の精度を低下させる大きな要因の一つとして、結合前対訳の曖昧性がある。これは、1 つの英語キーワードに対して、複数の他言語キーワードが対応付けられるという現象である。曖昧性の大きな英語キーワードを調査したところ、その多くは誤った対訳であり、さらにそれらの誤った対訳はほとんどが低頻度のものであることがわかった。そこで、曖昧性のある英語キーワードについて、その対訳を頻度順に並べ、頻度の高いもののみを残すという方法を提案する。【表４】に実験結果を示す。頻度の最も高い対訳のみを残した場合（Top-1）、獲得対訳数は約 16,000 対で、その精度は約 83% となり、やはり 32% から大きく向上することができた。

表 4：結合前フィルタリングによる獲得対訳数および精度の変化

フィルタ	日中対訳数	精度
なし	91,911	32%
Top-3	49,425	42%
Top-2	34,461	63%
Top-1	16,450	83%

4　おわりに

本研究では、二種類の言語で書かれた科学技術論文のキーワードリストから高精度に対訳対を獲得するために、キーワードリストのノイズに頑健な制約付き HMM を提案した。また、英語を中間言語として利用することで、直接対訳対を獲得することができない言語対においても対訳対を獲得し、さらにそれらの精度を向上するためのフィルタリング手法を提案した。

今後の課題としては、中英のキーワードリストデータを増やすことで、言語対によらず制約付き HMM が効果的であることを示すことと、中間言語を利用した対訳結合精度のさらなる向上が挙げられる。結合精度向上には、SVM に新たな素性を足すことや、SVM によるフィルタリングと、結合前フィルタリングを組み合わせることなどが考えられる。

参考文献

[1] Ren, F., Zhu, J. and Wang, H.: Web-based technical term translation pairs mining for patent document translation, International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE) 2010, pp. 1-8 (2010).

[2] Vogel, S., Ney, H. and Tillmann, C.: HMM-based word alignment in statistical translation, Proceedings of the 16th conference on Computational linguistics - Volume 2 , COLING '96, Stroudsburg, PA, USA, Association for Computational Linguistics, pp. 836-841 (1996).

[3] Cahyadi, D., Cromieres, F., Kurohashi, S.: Constrained Hidden Markov Model for Bilingual Keyword Pairs Alignment, Proceeding of 10th Workshop on Asian Language Resources, pp. 85-94 (2012).

[4] Och, F. J. and Ney, H.: A systematic comparison of various statistical alignment models, Computational Linguistic, Vol. 29, No. 1, pp. 19-51 (2003).

[5] Shezaf, D. and Rappoport, A.: Bilingual lexicon generation using nonaligned signatures, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10, Stroudsburg, PA, USA, Association for Computational Linguistics, pp. 98-107 (2010).

[6] Chu, C., Nakazawa, T. and Kurohashi, S.: Chinese Characters Mapping Table of Japanese, Traditional Chinese and Simplified Chinese, Proceedings of the Eighth Conference on International Language Resources and Evaluation (LREC'12), Istanbul, Turkey, pp. 2149-2152 (2012).