平成 23 年度 AAMT/Japio 特許翻訳研究会

報　告　書

機械翻訳及び辞書構築に関する研究
及び
海外調査

平成 24 年 3 月

一般財団法人 日本特許情報機構

目　　次

AAMT／Japio 特許翻訳研究会委員名簿

（敬称略・順不同）

委 員 長	辻井　潤一	マイクロソフトリサーチアジア・東京大学名誉教授・ マンチェスター大学客員教授・AAMT 前会長
副委員長	横山　晶一	山形大学大学院教授
〃	江原　暉将	山梨英和大学教授
委　　員	宮澤　信一郎	秀明大学教授
〃	梶　　博行	静岡大学教授
〃	黒橋　禎夫	京都大学大学院教授
〃	宇津呂　武仁	筑波大学大学院准教授
〃	二宮　崇	愛媛大学大学院准教授
〃	越前谷　博	北海学園大学准教授
〃	綱川　隆司	静岡大学助教
〃	範　　暁蓉	東京大学大学院　中川研究室
〃	安田　圭志	(独)情報通信研究機構
〃	熊野　明	東芝ソリューション(株)
〃	下畑　さより	沖電気工業(株)
〃	潮田　明	(株)富士通研究所
〃	三浦　貢	日本電気(株)
事 務 局	村上　嘉陽	AAMT/Japio 特許翻訳研究会東京事務局・(株)ナビックス
〃	河田　容英	〃　〃　〃
〃	高田　佳代子	〃　〃
オブザーバー	中川　裕志	東京大学大学院教授
〃	安藤　進	元多摩美術大学講師
〃	呉　先超	NTT コミュニケーション科学基礎研究所
〃	守屋　敏道	(財)日本特許情報機構
〃	森藤　淳志	〃
〃	藤城　享	〃
〃	大塩　只明	〃
〃	塙　　金治	〃
〃	三橋　朋晴	〃
〃	柿田　剛史	〃
〃	土屋　雅史	〃
〃	星山　直人	〃
〃	王　　向莉	〃

１．は　じ　め　に

マイクロソフトリサーチアジア　首席研究員
東京大学大学院情報理工学系研究科　名誉教授
AAMT/Japio 特許翻訳研究会委員長

辻井　潤一

長い研究の歴史を持つ機械翻訳であるが、ここ数年間、さまざまな応用場面での実用化が進んでいる。夢の技術とされた音声翻訳も、スマートフォンのアプリの一つとして使われるようになった。また、ウェブサーチの付属として、翻訳機能を提供することも普通になってきている。対象を特許に限っても、ヨーロッパの特許庁が、特許の翻訳を外部の企業と協力して本格的に行い始めたこと、アジアにおいても日本、中国、韓国の特許庁がそれぞれに機械翻訳の使用を本格化しようとしている。機械翻訳は、これらの試みを通して、研究機関での原理的な研究から、現実場面での使用を見据えた開発研究へと向かっている。

AAMT/Japio 特許翻訳研究会は、この現実場面での使用を見据えた機械翻訳の研究開発を促進するために、（１）機械翻訳システムの開発に従事する技術者だけでなく、（２）機械翻訳の原理的な研究を行っている大学や研究機関の研究者、また、（３）実際の特許の翻訳の工程を管理する機関の運営者、（４）翻訳業務にかかわる翻訳家など、背景の異なる人々に議論を深める場を提供している。また、公開の国際ワークショップやシンポジウムを企画することで、研究会の枠を超えて、特許翻訳の機械化に従事する人たちに連携の場を提供してきた。

本年度も、以上のような観点から活発な活動を行ってきた。8 回の研究会を開催し、翻訳評価の問題、専門用語の翻訳辞書の構築手法に関する問題、統計的機械翻訳と規則による翻訳システムの統合に関する問題など、特許翻訳の機械化の鍵となる課題を議論してきた。本報告書は、このような活動の成果を一般に公開するためのものである。また、本年度は、中国・厦門で開催された MT Summit に特許翻訳に特化したセッションや本会議に併設したワークショップを本研究会が中心となって運営するなど、国際的な連携でも成果を挙げた。本報告書には、この特別セッションと併設ワークショップの様子も含まれている。

本報告書が、知財の国際化に伴い、ますますその重要性を増している特許の多言語翻訳システムの開発、運用、利用に興味を持つ人たちの交流をさらに強めることに貢献できることを願っている。

２．１　対訳特許文を用いた同義対訳専門用語収集における推移的方式の評価

筑波大学大学院システム情報工学研究科
梁　冰，豊田　樹生，阿部　佑亮，
鈴木　敬文，宇津呂　武仁，山本　幹雄

2.1.1　はじめに

特許文書の翻訳は,他国への特許申請や特許文書の言語横断検索などといったサービスにおいて不可欠である．特許文書翻訳の過程において，専門用語の対訳辞書は重要な情報源であり，これまでに，対訳特許文書を情報源として，専門用語対訳対を自動獲得する手法の研究が行われてきた．森下らは，NTCIR-7 の特許翻訳タスクで配布された日英 180 万件の対訳特許文を用いて，対訳特許文からの専門用語対訳対獲得を行った[3]．この研究では，句に基づく統計的機械翻訳モデル[1]を用いることにより，対訳特許文から学習されたフレーズテーブル，要素合成法，Support Vector Machines (SVMs)[5]による機械学習を用いることによって，専門用語対訳対獲得を行った．しかし，森下らの手法では，ある日本語専門用語に対する英訳語を推定する際に,その日本語専門用語が出現する一つの対訳文に出現する英訳語のみを推定対象としているため,他の対訳文に出現している同義の専門用語対訳対を同定することができていない，という問題点があった．

そこで,先行研究[2]では,ある日本語専門用語が出現する複数の対訳文を入力として，同義の専門用語対訳対を同定する手法を提案する．提案手法では，対訳特許文および句に基づく統計的機械翻訳モデルのフレーズテーブルを用いて専門用語対訳対を収集し,それに対して，SVM を適用することにより，専門用語対訳対の同義・異義関係の判定を行う．この手法は，評価実験において，およそ 98%の適合率と 40%以上の F 値を実現した．

しかし，高い適合率に対して再現率が低いという問題点も見られた．そこで，本論文では，再現率の改善方法として，同義対訳専門用語の推移的同定の枠組みを提案する．この枠組みでは，SVM によって高適合率で同義と判定された専門用語対訳対を新たな中心的対訳対として選定し,それらの同義集合の和集合を元の中心的対訳対の同義集合として出力するという手順を再帰的に行う．この手法に対して行った評価実験の結果，95%の適合率と 32%の再現率を達成し，推移的同定の枠組みを適用しない場合と比べ，再現率が 4%向上した．さらに，推移的同定の枠組みにおいて，人手の介在を併用する場合では，95%以上の適合率と 50%以上の再現率を達成し，再現率をさらに 20%改善することができた．

表 1．作成された専門用語対訳対の同義候補集合中の対訳対数

	総要素数	134 個の集合の間の平均対数
同義候補集合 $\bigcup_{s_J} CBP(s_J)$	22,473	167.7
人手で同定した同義集合 $\bigcup_{s_{JE}} SBP(s_{JE})$	1,680	12.5

表 2. 同義判定の性能評価（%）

手法		適合率	再現率	F 値
ベースライン		67.0	54.3	68.0
SVM	適合率最大	**97.5**	28.7	43.9
	F 値最大	73.5	68.1	**70.5**

2.1.2 機械学習を用いた同義対訳専門用語の同定

2.1.2.1 適用手順

本論文では，先行研究[2]の場合と同様に，まず，表 1 に示すように，134 個の専門用語対訳対同義候補集合を生成した．そして，134 個の専門用語対訳対同義候補集合 $CBP(s_J)$ を全事例集合 CBP とし，互いに素な事例部分集合 $CBP_i(i = 1, \ldots, 10)$ に 10 分割する[1]．本論文では，機械学習のツールキットである TinySVM[2]を利用して，評価実験を行った．カーネル関数として，二次多項式カーネルを用いた．また，SVM の分離平面から，評価事例までの距離を信頼度とし，正例(すなわち，中心的対訳対と同義)判定に下限閾値を設定した．訓練の手順について，CBP_1，．．．，CBP_{10}の 10 個の部分集合のうち，8 個を訓練用事例集合として SVM の訓練を行い，残りのうちの 1 個を調整用事例集合として 2 種類のパラメータの調整を行い，最後の 1 個を評価用事例集合とした．以上の手順を 10 通り繰り返し，その平均値を算出し同義判定の性能評価を行った．なお，本論文で調整の対象としたパラメータは，SVM のソフトマージンを制約するパラメータ，および，分離平面から評価用事例までの距離の下限閾値である．

2.1.2.2 同義・異義判定のための素性

同義専門用語対訳対の同定に用いた素性は大きく，対訳対$\langle t_J, t_E \rangle$の特性を規定するものおよび，対訳対$\langle t_J, t_E \rangle$と中心的対訳対$\langle s_J, s_E \rangle$の間の関係を規定するものの 2 種類に分けられる．

2.1.2.3 評価結果

表 2 に，同義判定における性能の評価結果を示す．ベースラインとしては，「t_Jと s_Jが同一，または，t_Eと s_Eが同一」という条件を用いた．距離下限閾値およびソフトマージンのパラメータに対して，同義判定の適合率を最大化する調整を行った場合は，97.5%の適合率と 43.9%の F 値を達成した．一方，距離下限閾値およびソフトマージンのパラメータに対して，同義判定の F 値を最大化する調整を行った場合は，適合率 73.5%，適合率 68.1%，F 値 70.5%を達成した．

2.1.3 同義対訳専門用語の推移的同定

SVM(2節)による専門用語対訳対同義・異義自動同定の評価実験結果によって，適合率が高いものの，再現率が低い問題が存在していることが分かった．この問題を解決するため，SVMの同定結果に基づく推移的同定の枠組みを提案する．SVMにより高適合率で同定した同

---

[1] ただし，ここでは，134 個の中心的対訳対の集合を10 個に分割した．その際，各$CBPi(i = 1, \ldots, 10)$における正例(中心的対訳対と同義)・負例(中心的対訳対と異義) の数が，各$CBPi(i = 1, \ldots, 10)$の間で均等になるように，中心的対訳対の集合を分割した．

[2] http://chasen.org/~taku/software/TinySVM/

義集合は，中心的対訳対との同義同定が相対的に容易な事例の集合と考えられる．そこで，このような高適合率での同義同定を漸進的に行うことにより，中心的対訳対との同義同定を直接行うことが困難な事例を同定することができ，同義同定の再現率の改善につながるというのが，この枠組みの基本的な考え方である．

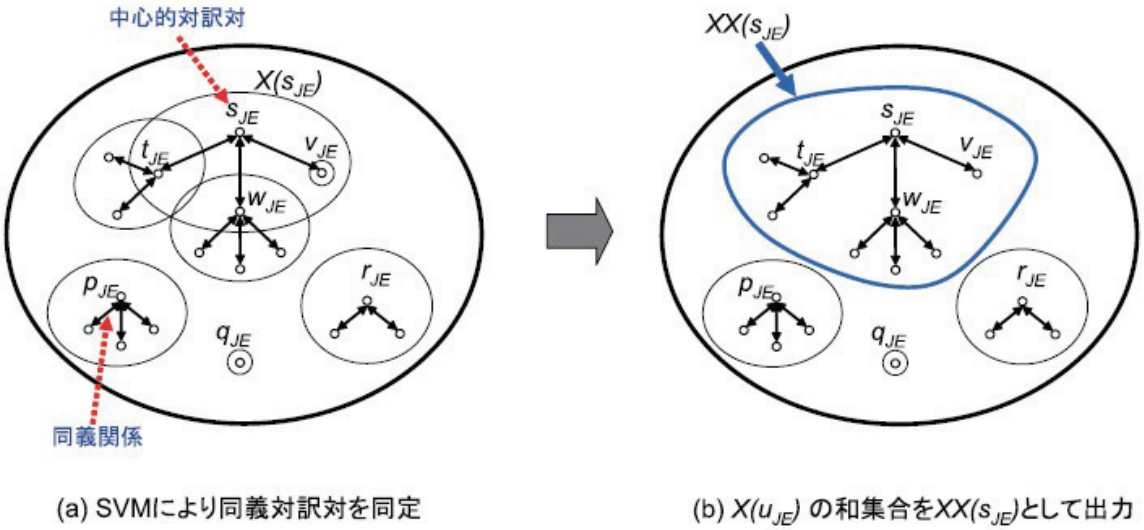


図 1. 同義対訳専門用語の推移的同定手順（人手の介在を併用しない）

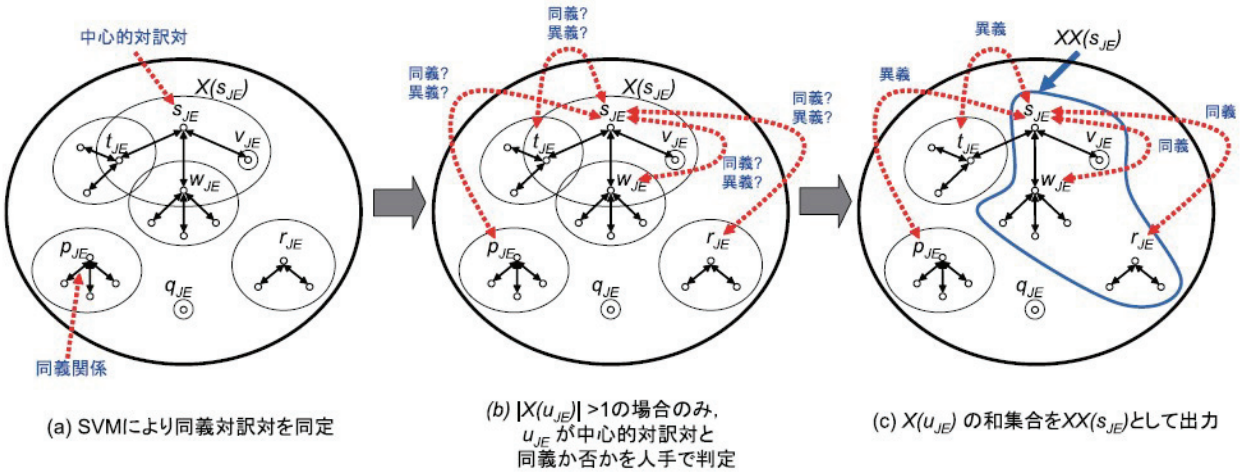


図 2. 同義対訳専門用語の推移的同定手順（人手の介在を併用）

2.1.3.1 推移的同定の手順

以下では，同義専門用語対訳対の推移的同定の手順を述べる．

ステップ1 専門用語対訳対の同義候補集合$CBP(s_{JE})$の要素に対して，あらゆる$u_{JE}=\langle u_J, u_E\rangle$と$v_{JE}=\langle v_J, v_E\rangle$の組（ただし，$u_{JE} \neq v_{JE}$）を作成し，それらの組にSVM(2節)を適用し，同義・異義関係を判定する．

ステップ2 それぞれの$u_{JE}=\langle u_J, u_E\rangle$（$\in CBP(s_{JE})$）に対し，$u_{JE}=\langle u_J, u_E\rangle$と同義の$v_{JE}=$

〈v_J, v_E〉(∈ $CBP(s_{JE})$) (≠ u_{JE})を集合$X(u_{JE})$の要素とする(図1(a)，図2(a))[3].

$$X(u_{JE}) = \left\{ v_{JE} = \langle v_J, v_E \rangle (\in CBP(s_J)) \middle| \begin{array}{l} v_{JE} = u_{JE} \text{,または ,SVM(2節)により} \\ v_{JE} \text{と } u_{JE} \text{を同義であると判定} \end{array} \right\}$$

ステップ3 このステップは，人手の介在を併用するか否かにより，以下の2つの方式に分けられる．

人手の介在を併用しない推移的同定 人手の介在を併用しない推移的同定は,複数の中心的対訳対間の同義・異義関係を判定する際，SVMによる自動同定結果を利用する方式である．SVM(2節)により中心的対訳対s_{JE}と同義であると判定された専門用語対訳対u_{JE}の集合を$SBP'(u_{JE})$と定義する．

$$SBP'(s_{JE}) = \left\{ u_{JE} = \langle u_J, u_E \rangle (\in CBP(s_J)) \middle| \text{SVM(2節)により } u_{JE} \text{と } s_{JE} \text{を同義であると判定} \right\}$$

また，SVMにより，中心的対訳対s_{JE}と同義であると判定した専門用語対訳対u_{JE}に対して，ステップ2で定義した$X(u_{JE})$の和集合を$XX(s_{JE})$と定義する(図1(b))．

$$XX(s_{JE}) = \bigcup_{u_{JE} \in SBP'(s_{JE})} X(u_{JE})$$

この方式では，$XX(s_{JE})$ を中心的対訳対s_{JE}の同義対訳専門用語集合として出力する．

人手の介在を併用した推移的同定 人手の介在を併用した推移的同定は，複数の中心的対訳対間の同義･異義関係を判定するとき,人手による判定を利用する方式である．それぞれの専門用語対訳対u_{JE}= 〈u_J, u_E〉(∈ $CBP(s_{JE})$)に対し，$X(u_{JE})$ ＞1の場合のみ[4]，u_{JE}は中心的対訳対s_{JE}と同義であるか否か(すなわち，u_{JE} ∈ $SBP(s_{JE})$)を人手で判定する(図2(b))．
また，人手により，中心的対訳対s_{JE}と同義であると判定した専門用語対訳対u_{JE}に対して，ステップ2で定義した$X(u_{JE})$の和集合を$XX(s_{JE})$と定義する(図2(c))．

$$XX(s_{JE}) = \bigcup_{\substack{u_{JE} \in SBP(s_{JE}) \\ |X(u_{JE})|>1}} X(u_{JE})$$

この方式では，$XX(s_{JE})$を中心的対訳対s_{JE}の同義対訳専門用語集合として出力する．

---

[3] ここで，v^1_{JE} およびv^2_{JE} のいずれも，SVMにより，u_{JE} と同義であると判定され，その一方で，v^1_{JE} とv^2_{JE} は異義であると判定される場合は，本論文では，v^1_{JE} とv^2_{JE} の両方を$X(v_{JE})$の要素とする．

[4] この条件は，SVMが少なくとも一つの専門用語対訳対v_{JE}がu_{JE}と同義であると判定する場合に相当する．この条件が成り立たない場合は，u_{JE}が中心的対訳対s_{JE}と同義であるか否かの人手による判定を行わない．

表 3. 同義対訳専門用語の推移的同定の評価結果（%）

調整用事例における適合率の条件	適合率 / 再現率 / F値		
	推移的同定なし	推移的同定あり	
		人手の介在を併用しない	人手の介在を併用
> 80%	79.3 / 53.9 / 63.6	78.4 / **59.1** / **66.6**	81.3 / **89.9** / **85.1**
> 85%	85.1 / 46.4 / 59.7	84.2 / **49.6** / **61.6**	86.9 / **80.9** / **83.4**
> 90%	89.0 / 38.6 / 53.3	89.7 / **42.7** / **57.5**	91.3 / **69.1** / **78.2**
> 95%	94.1 / 27.6 / 42.4	95.2 / **32.1** / **47.9**	95.2 / **53.1** / **67.9**

2.1.3.2 評価結果

本論文では，複数の中心的対訳対間の同義・異義関係を判定し，複数の同義対訳専門用語集合を統合することにより，同義同定の再現率を改善するという推移的同定の枠組みのもとで，評価実験を行った．具体的には，3.1節で述べた方式を評価した．

2.1節で述べたように，調整用事例集合を用いて，距離下限閾値を調整することにより，判定結果の適合率を変化させることができる．評価実験において，調整用事例における判定結果の適合率が80%以上，85%以上，90%以上，95%以上のときのそれぞれの距離下限値を利用した場合の評価用事例における評価結果を表3に示す[5]．

全体として，人手の介在を併用なしの推移的同定の評価結果においては、推移的同定なしのときの評価結果と比べ，再現率を平均4%以上改善した．さらに，人手の介在を併用した推移的同定の評価結果においては，人手の介在を併用しない推移的同定の評価結果と比べ，適合率は平均2%以上向上し，再現率はさらに平均30%改善された．しかし，人手の介在を併用しない推移的同定方式においては，再現率の増加は一サイクル目の推移的同定において最大となり（表3の評価結果に示した結果），それ以降のサイクルにおいて，再現率をさらに改善することができなかった．一方，人手の介在を併用した推移的同定方式は，高い適合率を保ちながら，再現率を大幅に改善した．言い換えると，この再現率は推移的同定という枠組みの現段階における再現率の上限値であるといえる．

2.1.4 関連研究

文献[4]は，対訳専門用語の同義判定に機械学習を用いており，手法の点においても，また，機械学習で用いている素性の点においても，本論文の手法と密接に関連している．しかし，文献[4]では，同義判定の対象とする対訳専門用語の収集を手動で行っており，手法の適用範囲が非常に限定されている．一方，本論文の手法は，毎年に公開される対訳特許テキストから，同義判定の対象とする対訳専門用語の収集を自動で行っており，文献[4]と比較して，手法の適用範囲が限定されないという点で，優れていると言える．

[5] 参考として，各参照用専門用語対訳対の同義集合$SBP(s_{JE})$ の要素u_{JE} のうち，$|X(u_{JE})| = 1$の平均要素数を測定した．この数は，実際，どのくらいの要素u_{JE} に対して，中心的対訳対s_{JE} と同義であるか否かの判定を行う必要がないかを表す．一つの中心的対訳対あたりの参照用同義対訳専門用語の数は12.5個であるが，表3に示す結果においては，この数は，それぞれ，“> 80%”の場合は0.9，“> 5%”の場合は1.4，“> 90%”の場合は2.2，“> 95%”の場合は4.0となった．

2.1.5 おわりに

本論文では,同義対訳専門用語の自動同定において再現率が低いという問題点を改善するため，推移的同定の枠組みを構築した．評価実験において，95%以上の適合率と32%の再現率を達成し，再現率を4%改善した．さらに，新たな中心的対訳対を選定する際に，人手の介在を併用した場合では，95%以上の適合率と50%以上の再現率を達成し，再現率をさらに20%改善した．今後の課題としては，中心的対訳対の同義候補集合の生成(文献[2]を参照)の過程を再帰的に行う方式を開発することが重要であると考えられる．

参考文献

[1] P.Koehn, H.Hoang, A.Birch, C.Callison-Burch, M.Federico, N.Bertoldi, B.Cowan, W.Shen, C.Moran, R.Zens, C.Dyer, O.Bojar, A.Constantin, and E.Herbst. Moses:Open source toolkit for statistical machine translation. In *Proc. 45th ACL, Companion Volume*, pp.177-180, 2007.

[2] 梁冰，宇津呂武仁，山本幹雄．対訳特許文を用いた同義対訳専門用語の同定と収集．言語処理学会第17 回年次大会論文集，pp. 963-966，March 2011.

[3] 森下洋平，梁冰，宇津呂武仁，山本幹雄．フレーズテーブルおよび既存対訳辞書を用いた専門用語の訳語推定．電子情報通信学会論文誌，Vol. J93-D，No. 11，pp. 2525-2537，2010.

[4] T.Tsunakawa and J.Tsujii. Bilingual synonym identification with spelling variations. In *Proc. 3rd IJCNLP*, pp.457-464, 2008.

[5] V.N.Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

２．２　語学学習サイトウェブページからの対訳語抽出

東京大学　　範　暁蓉
愛媛大学　　二宮　崇

2.2.1　はじめに

対訳語辞書は機械翻訳や多言語横断検索システムなどのための非常に重要な言語リソースとなっている。対訳語自動抽出の研究は今までに様々な手法が提案されており、抽出元となるリソースによって、既存の対訳辞書からの対訳語抽出、対訳コーパスからの対訳語抽出、単一言語コーパスからの対訳語抽出などに大きく分けられる。この中で、対訳コーパスからの対訳語抽出手法はもっとも長く研究されており、精度も高い。しかし、この手法において十分な量の対訳語辞書が得られるかどうかは対訳コーパスの量に大きく依存しており、既存の対訳コーパスでは量的にまだ不十分である。特に、大規模な日中コーパスは少なく、文書の内容も限られている。

本稿では、中国の日本語学習者に向けの語学学習サイトウェブページから、大規模な日中対訳コーパスを構築することを行う。構築された対訳コーパスから対訳語自動抽出手法を適用して、構築されたコーパスの性能を考察する。

本稿の構成は以下のようになっている。2.2.2 節では、語学学習サイトウェブページから日中対訳コーパスを構築する手法を説明する。2.2.3 節では、語学学習サイトウェブページからの対訳コーパス構築の手順と結果を報告する。2.2.4 節では、構築された対訳コーパスから得られる対訳語抽出の実験について説明する。2.2.5 節で本稿の主旨をまとめ、今後の課題について述べる。

2.2.2　日中対訳コーパス

対訳コーパスは対訳語自動抽出研究において、初めて利用された言語リソースのひとつであり、もっとも有効なリソースである（Brown et al. 1990)。日中対訳コーパスを生成するには大量の日本語文と中国語文が必要であり、対訳文を人手で作成するには膨大な時間と労力を必要とする。たとえば、北京日本学研究センターが 2003 年度公開された「中日対訳コーパス」（徐一平ら、2002）の開発には約 3 年の時間を要した。

2.2.2.1　ウェブ上多言語文書

ウェブにはボランティア翻訳者によって翻訳された文書が大量に存在する。この翻訳文と原文書を収集し、対訳コーパスは生成できる。石坂達也ら（2009）はこの手法で、大規模な日英対訳コーパスを生成し、一般に公開されている。しかし、このように自動生成された日中対訳コーパスは量的にまだ少ない。その原因は日本語と中国語の訳文を入手することが難しいことにある。

ウェブ上には翻訳文と原文から成る多言語文書がたくさんあり、この中で、Wikipedia は最もよく使われているリソースである。Wikipedia 上の中国語資源は二種類にわけられる。一つは、原文が中国語となっている文書である。このタイプの文書は日本語訳文が少ないという問題点がある。もう一つは原文が英語あるいは日本語と中国語以外の言語で、いくつかの言語に翻訳され

表 1　日本語と中国語の対訳対応が悪い例

言語	文
英語	The **electron** (symbol: e−) is a subatomic particle with a negative elementary electric charge.
日本語	**電子**（でんし、英語: Electron）とは、宇宙を構成する素粒子のうちのレプトンの 1 つである。素粒子の標準模型では、第 1 世代の荷電レプトンとして位置づけられる。
中国語	**电子**（Electron）是一种带有负电的亚原子粒子，通常标记为 e^-

た文書である。後者のタイプの文書については、日本語と中国語の対訳の対応が良くないという問題点がある。表 1 に例を示す。表 1 は、Wikipedia 上の「Electron」に関する英語と日本語と中国語の説明の一部である。この例における対訳関係は、英語と中国語訳が対応し、日本語は英語と中国語両方とも対応しないという状況になっている。このようなウェブページから、日中の対訳コーパスを構築することは難しい。

2.2.2.2　語学学習サイト

国際交流基金によると[1]、現在、中国の日本語学習者の数は約 83 万人で、その増加は著しく、日本語能力試験海外受験者数は世界で最も多い。このたくさんの日本語学習者のために多くの日本語学習サイトが作られている。自然な日本語を勉強するため、毎日、日本語学習サイトはいくつかの日本語文書と中国語翻訳文を提供している。日本語学習サイトが提供する日本語とその翻訳文は三つの特徴がある。

（1） 図 1 に示されるように、日本語文章は文単位で中国語に直訳されている。このような文から対訳コーパスを作成することは容易である。

（2） 日本語学習者の理解を容易にするために、文章の重要な日本語単語と中国語訳語も提供している。図 2 の上の部分がこれを示した。これらの対訳単語対は対訳語の正例として用いられる。

（3） 図 2 の下の部分に、この文章の支持度を示した。これは中国語翻訳文の品質の判断基準として使える。

このサイトから 50 篇の日本語文章と翻訳文を収集し、人手で分析すると、98%の文が中国語と日本語の間で対応関係にあった。このため、語学学習サイトから日中対訳コーパスの構築が可能であると考える。

[1]「海外日本語教育機関調査」: 国際交流基金（ジャパンファウンデーション）が、各国の在外公館、財団法人交流協会の協力を得て、海外の日本語教育機関を対象に、学習者数、教師数、学習目的、問題点などを問うために実施しているアンケート調査。

中国語でも「餅(ビン)」という食べものがあり、日本の「餅(もち)」と漢字が同じですが、実体は全く異なります。(中国語の「餅(ビン)」は、小麦粉を用いて火を通した、平たく丸い食品を言います。)

汉语里，有称做餅(ビン)这样的食物，虽然与日本的"餅(もち)"汉字相同，但是实质完全不一样。(汉语里，将小麦粉做的用火加热而成的扁平的圆圆的食品叫做"餅(ビン)"。)

日本のお餅は、特につきたては、弾力性があります。女性の美しく弾力性のある肌をほめる表現で、「もち肌」と日本では言うのはそのためです。またお雑煮、お汁粉の写真を見ていただくとわかるかもしれませんが、お餅には粘着性があります。そのため、お餅をあわてて食べると、のどにくっついて、詰まって窒息死す るおそれがあります。特に高齢者には注意が必要です。もし中国の人が日本でお餅を初めて食べる時には、この点に十分注意してください。

日本的年糕，特别是刚捣好的时候很有弹性。因此在日本，"饼肌"是用来赞美女性拥有美丽弹性肌肤的表达方式。另外如果看了年糕汤，小豆汤的图片后，会发现年糕具有粘着性。因此，急匆匆地吃年糕的话，会粘在咽喉里堵塞着，恐怕会导致窒息死亡。特别是老年人应当注意。如果中国人在日本第一次吃年糕的话，请谨记这一点。

図 1　日本語学習サイトの対訳文の例

図 2　重要な日本語単語の中国語単語訳

2.2.3　語学学習サイトから日中対訳コーパスの構築

以下の手順で対訳コーパスを構築する。

（1）　対訳文書の収集

（2）　文書の整形

（3）　文の対応付け

表２　タグの例

中国語でも「餅（ビン）」という食べものがあり、日本の「餅（もち）」と漢字が同じですが、実体は全く異なります。（中国語の「餅（ビン）」は、小麦粉を用いて火を通した、平たく丸い食品を言います。）

汉语里,有称做餅(ビン)这样的食物,虽然与日本的“餅(もち)”汉字相同，但是实质完全不一样。（汉语里，将小麦粉做的用火加热而成的扁平的圆圆的食品叫做“餅（ビン）”。）

2.2.3.1　対訳文書の収集

学習サイトから文書を収集するためのクローラを作成した。HTML のリンクをたどって文書を収集し、一週間に一回程度更新する。収集された文書には次の 3 種類がある。

（１）全日本語文書
（２）全中国語文書
（３）日本語と中国語を含む文書。

（１）と（２）は対訳文書となっていないため、収集しない。（３）はさらに 3 種類にわけられる。

（ア）日本語文書の中心内容だけが中国語に翻訳されている。
（イ）翻訳文章がなくて、関連中国語文書がある。
（ウ）本当の日中翻訳文がある。

（ア）と（イ）は収集せず、（ウ）だけ収集した。

2.2.3.2　対訳文書の整形

収集した文書には原文と翻訳関係にない中国語の説明があったり、文の途中に改行があったり、日本語の発音も含まれている。収集した文書を整形せずに文の対応付けを行うと、対応付けの精度が低下する。よって、収集した文書を１行１文になるように整形する。

表 2 はタグ付きの収集された文書を示している。収集した文書の html タグを分析して、日本語原文と中国語翻訳語のタグが分かる。 タグの内容が日本語原文、 の内容が中国語翻訳文である。この二つのタグの内容を取り出して、対訳文書ができる。

2.2.3.3　文の対応付け

表 3　文の完結を表す終止符

。　！　？　…　：

表 4　対訳文の例

日本語文	中国語文
強い寒気の影響で昨日から広い範囲で雪が降り、北陸や近畿北部では局地的に大雪になりました。	受强冷空气影响昨天日本大范围内出现降雪，北陆和近畿地区局部地区大雪。
また、低気圧が東北北部を通過したため、今朝は北日本で局地的に風が強まりました。	此外，由于低气压经过东北北部，今晨日本北部局部地区风力加强。
生活情報番組「発掘!あるある大事典」の制作費は 3250 万円。	生活情报节目《发掘！あるある大事典》的制作费为 3250 万日元。

日中辞書の入手が困難であって、学習サイトから収集した対訳文書のレベルの対応付けは句点などの終止符で判断する。学習サイトの文書はおおよそ直訳になっているため、この手法で高い精度の文対応を作ることができる。本研究では、表 3 の終止符が文の完結とみなす記号となる。

2.2.3.4　日中対訳コーパス収集結果

今回の実験では、中国の最も有名な語学学習サイト「沪江日语」[2]を利用する。生成した対訳文の例を表 4 に示す。今まで、収集された対訳語の数は 139,790 である。

2.2.4　対訳語抽出実験

学習サイトから生成した日中対訳コーパスから、複合語対訳語抽出実験を行った。抽出手法は Fan (2009) の手法を使用する。この手法を大まかに説明する。

日本語文は J，中国文は C とする。Wj はJの含まれる単単語，Wc は C に含まれる単単語である。Pj は J に含まれる複合語、Pc は C に含まれる複合語。文に含まれる単語は形態素解析により与えられる。文に含まれる複合語は用語抽出により与えられる。

2.2.4.1　対訳語対訳確率計算

まず、日本語文、中国語文それぞれで形態素解析を行う。形態素解析の結果に対し、アラインメントを行う。Wj と Wc の対訳確率 P(Wj,Wc) ができる。

次に、日本語文、中国語文それぞれで用語抽出を行う。形態素解析の結果は用語抽出された複合語により修正される。文の中に複合語があれば、複合語はひとつのアラインメント単位になる。

次に、修正された結果でアラインメントを行う。Pj と Pc の対訳確率 P(Pj,Pc) ができる。

複合語に含まれる単単語間の対訳確率による複合語の対訳確率は式 1 で定義する。

[2] http://jp.hujiang.com/

表 5　複合語の数

言語	語数
日本語複合語	49,612
中国語複合語	33,747
複合語対訳語	23,518

表 6　抽出結果の例

日本語	中国語訳	日本語	中国語訳
NHK 連続テレビ小説	NHK 连续剧小说	京漬物	京都渍菜
女子大生	女大学生	中国民俗通史	中国民俗通史
平安時代	平安时代	重要無形文化財	重要无形文化财产
修学旅行	修学旅行	認知意味論	认知语义学
掲示板	贴吧	質問状	提问书

$$assoc(\mathrm{Pj},\mathrm{Pc}) = \frac{\sum_{k,m} assoc(\mathrm{Wj}_k, \mathrm{Wc}_m)}{\max(length(\mathrm{Pj}), length(\mathrm{Pc}))} \qquad (1)$$

Wj_k は Pj に含まれる単語で、Wc_m は Pc に含まれる単語である。*length*(Pj) は Pj に含まれる単単語の数で、*length*(Pc)は Pc に含まれる単語の数である。
最後に、*sim*(Pj,Pc)で Pj と Pc の対訳確率が算出される。

$$sim(S,T) = w_1 align(S,T) + w_2 assoc(S,T) \quad \text{s.t.} \quad w_1 + w_2 = 1 \qquad (2)$$

2.2.4.2　実験

日本語の形態素解析を茶筅で行って、中国語の形態素解析を ICTCLAS で行った。GIZA++でアラインメントを行った。

抽出された複合語と対訳語の数を表 5 に示す。表 6 は抽出の結果の一部を示す。語学学習サイトが提供する日本語文は主に日本に関する歴史や、風土と人情や、今流行する映画とテレビドラマおよび一番人気な掲示板の内容などである。これらに関連する対訳が抽出されていることがわかる。

2.2.5 まとめ

本稿では語学学習サイトウェブページ上の日中対訳文書を収集し、文対応日中対訳コーパスを構築した。また、構築した対訳コーパスを用いて対訳語抽出実験を行った。実験結果より、語学学習サイトウェブページから構築された対訳コーパスは、対訳語抽出のための有用なリソースであることがわかった。しかし、今回の日中対訳コーパスを構築するとき、日中対訳辞書を用いなかったため、文の対応付け処理が十分ではなかった。今後は文対応の質をあげるための文対応づけの研究を行いたいと考えている。

参考文献

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vncent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer and Paul S. Roossin: A Statistical Approach to Machine Translation. Computational Linguistics, 16(2), pages. 79-85, 1990.

徐 一平, 曹 大峰: 中日対訳語料庫的研製与応用研究論文集, 2002.

Tatsuya Ishisaka, Masao Utiyama, Eiichiro Sumita and kazuhide Yamamoto: Development of a Japanese-English Software Manual Parallel Corpus, In Proceedings of The Machine Translation Summit XII, 2009.

Xiaorong Fan, Nobuyuki Shimizu, and Hiroshi Nakagawa: Automatic extraction of bilingual terms from a Chinese-Japanese parallel corpus. In Proceedings of the 3rd International Universal Communication Symposium (IUCS '09), pages 41–45, 2009.

２．３　コンパラブルコーパスを用いた要素合成法によるターム翻訳の改良

静岡大学情報学部　梶　博行
綱川隆司
小松原慶啓

【要旨】コンパラブルコーパスを用いた要素合成法によるターム翻訳の改良を提案する。対応づけられた文書対からなる２言語コーパスから、タームに含まれる単語列の対訳とその相関値から構成される対訳辞書を獲得する。そして、この辞書を参照して、入力タームに対する合成訳語を確信度スコア付きで生成する。このようにして、入力タームに対し、できるだけ多くの訳語候補の中から正しい訳語を選択することができる。日英の科学技術文献抄録からなるコンパラブルコーパスを用いた実験を行い、コーパスから獲得した相関付き対訳辞書を用いた要素合成法が通常の対訳辞書を用いた要素合成法より高い性能をもつことを実証した。今後の課題として、相関付き対訳辞書の逐次的改良方法、確信度スコアの精密化、語順の変化を許す合成翻訳モデルの拡張があげられる。

2.3.1　はじめに

テクニカルタームの翻訳は文書翻訳や言語横断情報検索における重要な課題である。一つの専門分野のタームをすべてカバーする対訳辞書が存在しないことは明らかである。しかしながら、テクニカルタームの多くは複合語であり、いくつかの専門分野における日本語のテクニカルタームの８８％は合成的な英語訳をもつ (Tonoike, et al. 2006)。したがって、テクニカルタームの翻訳において要素合成法が重要な役割を果たすといえる。

要素合成法の性能は当然のことながら参照する対訳辞書に依存する。タームの構成要素に対する適切な訳語を辞書が与えなければ、正しい訳語を生成することはできない。同時に、構成要素の各々に対しできる限り多くの訳語を辞書が与えるとき、合成的に生成される多くの訳語候補の中から正しい訳語を選択することは困難である。前者の問題を解決するため対訳辞書のカバー率を向上させると後者の問題がいっそう深刻になることに注意する必要がある。

本稿では、２言語コーパスを用いた要素合成法の改良を提案する。すなわち、２言語の単語列の対とそれらの間の相関値からなるカバー率の高い対訳辞書を２言語コーパスから獲得する。そして、入力タームに対し、構成要素の単語とその訳語の間の相関値に基づく確信度スコアとともに合成的に訳語候補を生成することにより、ランク付き訳語候補リストを生成する。本提案の新規性は、２言語コーパスからの対訳辞書の獲得ではなく、確信度スコア付きの改良された要素合成法にある。

本稿で提案するフレームワークはパラレルコーパスとコンパラブルコーパスの両方に適用することができる。一般にパラレルコーパスのほうがコンパラブルコーパスより信頼度の高い相関値の付いた対訳辞書を生成することができる (Och and Ney 2003; Koehn et al. 2003)。しかしながら、大規模なパラレルコーパスが利用できる分野はほとんどない。したがって、本稿では、入力コー

パスとしてコンパラブルコーパス、より具体的にいうと対応づけられた文書対からなるコーパスを想定する。より多くの分野で利用可能であるが、より信頼度の低い相関値の付いた対訳辞書しか生成できないと思われる疎なコンパラブルコーパスの利用は本稿の範囲外である (Fung and Yee 1998; Rapp 1999; Andrade et al. 2010; Ismail and Manandhar 2010; Morin and Prochasson 2011)。

パラレルコーパスやコンパラブルコーパスからの対訳辞書獲得に関しては多くの研究があるが、そこでのタスクは、通常、入力コーパスに含まれるタームの対訳を抽出することである。対訳辞書獲得方法は、通常、入力コーパス中に出現するソース言語のタームに対して獲得されるターゲット言語の訳語の再現率と適合率で評価されてきた (Fung and Yee 1998; Rapp 1999; Cao and Li 2002; Tanaka 2002)。これに対し、本稿でのタスクは入力コーパスに出現しなくてもタームを翻訳することである。したがって、入力コーパスとは独立に用意した入力タームのテストセットに対して生成される訳語の精度によって、提案するフレームワークを評価する。文書翻訳や言語横断情報検索といった対訳辞書の実際の応用を考えたとき、このタスク設定は自然である。

2.3.2 課題と提案するフレームワーク

日本語のターム“光通信”とその英語の訳語“optical communication”の組を考えてみよう。人間は“光”と“optical”が対応し、“通信”と“communication”が対応していると認識することができる。言い換えると、“光通信”から“optical communication”への翻訳は合成的である。しかし、電子化された日英対訳辞書のほとんどは日本語の名詞（例：“光”）と英語の形容詞（例：“optical”）の対応を含んではいない。そのため、自動的な要素合成法では、通常、入力ターム“光通信”に対して正しい訳語“optical communication”を生成することができないのである。

日本語の名詞“光”と英語の形容詞“optical”の組が対訳辞書に登録されたとしよう、対訳辞書は、“光”に対し“optical”だけでなく“light”、“ray”、“beam”など多くの可能な訳語を与えるであろう。同様に、“通信”に対し“communication”、“correspondence”、“report”など可能な訳語を与えるであろう。そのため、要素合成法は“optical communication”、“optical correspondence”、“optical report”、“light communication”、“light correspondence”など多数の訳語候補を生成し、その中から正しい訳語を選択しなければならない。

上の例のように、要素合成法によるタームの翻訳には二つの問題がある。不完全な対訳辞書とほとんどが誤りの多数の訳語候補である。これらの問題を解決するため、本稿では、(1) 2 言語コーパスからの相関付き対訳辞書の獲得と(2)確信度スコア付きの合成訳語の生成という二つのステップからなるフレームワークを提案する。

(1) 2 言語コーパスからの相関付き対訳辞書の獲得

相互に関連する文書の組からなるコンパラブルコーパスが利用可能であるとし、対応する文の組における共起統計に基づく、二つの言語の語の間の相関を計算する方法を利用する (Matsumoto and Utsuro 2000)。この方法はもともとパラレルコーパスへの適用を意図したものであるが、文書の組を文の組のように扱うことによりコンパラブルコーパスに適用することができる (Utsuro et al. 2003)。個々の文書が小さい限り動作可能と思われる。この方法の利点は、コンパラブルコーパ

スに適用可能な他の対訳獲得方法と違って、種となる対訳辞書を必要としないことである。

ここでの目的は実際のタームではなくタームの構成要素に対する高カバー率の対訳辞書を作ることである。構成要素の間の対応の多くは、“光”と“optical”のような単純語間の対応であるが、“薄膜”と“thin film”のような単純語と複合語の対応、逆に“移動 体”と“mobile”のような複合語と単純語の対応もある。したがって、単純語の組だけでなく単純語と複合語が混じった組も抽出することが必要である。しかしながら、複合語を同定することは必ずしも容易でない。また、実際的な立場からは、タームに含まれる任意の単語列に対して可能な訳語を与えるような対訳辞書が望ましい。長い単語列の対訳の組が与えられると、タームに対し正しい訳語が生成される可能性が高くなると思われるからである。したがって、タームに含まれる任意の単語列をタームの構成要素と考え、ソース言語の単語列とターゲット言語の単語列の間の相関を計算することとする。

(2) 確信度スコア付きの合成訳語の生成

合成的に生成される多数の訳語候補の中から正しい訳語を選択するため、訳語候補の各々に対し確信度スコアを計算することとする。構成要素の対訳は相関値とともに獲得されていることに注意されたい。この相関を構成要素の訳語の確信度スコアとみなし、合成訳語の確信度スコアを構成要素の訳語のスコアに基づいて定義する。

ステップ 1 で述べたように、対訳辞書は単語に対するだけでなく単語列に対する訳語を与えている。しかし、その相関値すなわち確信度スコアはあまり信頼度が高くない。それゆえ、単語列に対して対訳辞書が与える訳語を再評価することとする。すなわち、単語列が対訳辞書に含まれていても、それに対する合成訳語を生成し、対訳辞書が与える確信度スコアと合成的に計算される確信度スコアを組み合わせる。

以下の二つの節で、提案するフレームワークの二つのステップを詳細に述べる。そこでは、ソース言語、ターゲット言語をそれぞれ日本語、英語とするが、形態素の扱いなど言語固有の事項について修正すれは、提案するフレームワークは任意の言語対に適用することができる。

2.3.3 要素合成法のための対訳辞書の獲得

日本語文書と英語文書の両方からタームに含まれる単語列をすべて抽出する。日本語のタームの多くは<名詞>+、すなわち 1 個以上の名詞の列であり、英語のタームの多くは<形容詞>*<名詞>+、すなわち 0 個以上の形容詞の列に続く 1 個以上の名詞の列である。ここに、形容詞には動詞の現在分詞形や過去分詞形も含まれる。現在のところ、前置詞句を含むタームなど、より複雑な構造をもったタームは取り扱わない。したがって、日本語では名詞の列、英語では名詞と形容詞の列を抽出する。

日本語の単語列 J と英語の単語列 E の相関を Dice 係数を用いて定義する。すなわち、

$$C(J,E)=\frac{2\cdot g(J,E)}{f(J)+f(E)}, \quad [1]$$

ここに、$f(J)$と$f(E)$はそれぞれJが生起する日本語文書の数とEが生起する英語文書の数である。また、$g(J,E)$はJとEが共起する日本語と英語の文書の組の数である。

一つの文書中の単語列の生起頻度は無視する。その理由は、提案するフレームワークは非パラレルコーパスへの適用を意図しており、その場合、日本語文書における単語列の生起頻度と対応する英語文書における対応する英語の単語列の生起頻度が同程度であるわけではないからである。また、（移動 体, mobile）や（薄膜, thin film）の例のように単語列の長さが言語間で保存されるとは限らないので、単語列の長さも無視する。

最大単語列、すなわちより長い単語列の部分列でない単語列、と非最大単語列を区別することを述べておくことが必要である。対応する日本語と英語の文書の組において、日本語の最大単語列、非最大単語列は英語の最大単語列、非最大単語列とそれぞれ対応する傾向がある。したがって、ある文書対に共起する最大単語列の組あるいは非最大単語列の組に対しその文書対を 1.0 とカウントするのに対し、ある文書対に共起する最大単語列と非最大単語列の組に対しその文書対を 0.5 とカウントする。対応づけられた文書対に“光通信”と“optical communication”がともに最大単語列として生起すると仮定する。この文書対は（光通信, optical communication）、（光, optical）（通信, communication）に対して 1.0 とカウントされるが（（光, communication）、（通信, optical）に対しても 1.0 とカウントされることに注意）、（光, optical communication）、（通信, optical communication）、（光通信, optical）、（光通信, communication）に対しては 0.5 とカウントされる。このようにして、複合語とその構成要素の間の混乱を軽減する。

入力コーパス中に低頻度で出現する単語列に対しては相関値の信頼度は低いので、単語列が出現する文書数に対する閾値θ_fを設定する。そしてθ_f以上の文書に出現する日本語単語列と英語単語列の全ての組に対して相関を計算する。日本語タームを英語に翻訳することを意図しているので、日本語単語列の各々に対し、相関値の降順に上位 N_1 個の英語単語列を選択する。（第５節で述べる実験では、θ_fを 10、N_1 を 20 に設定した。）

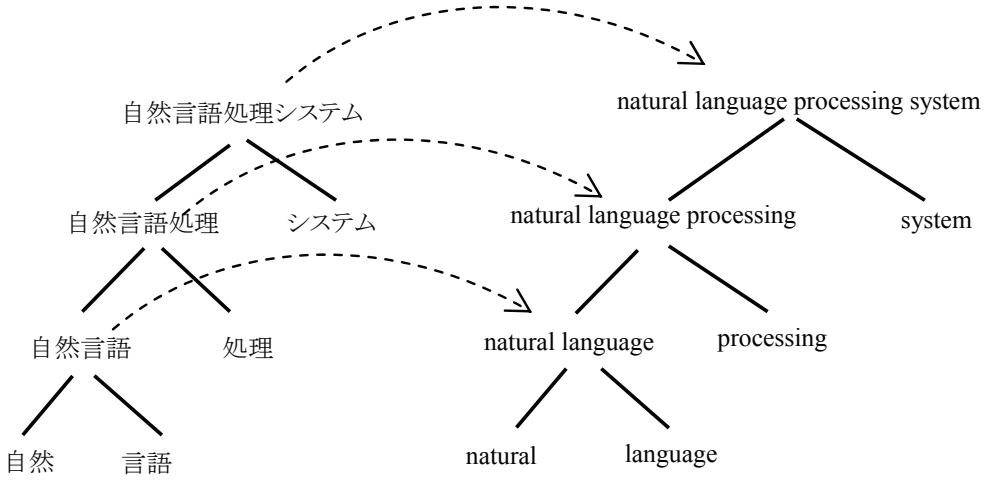
2.3.4 確信度スコア付きの要素合成法

タームは、図１に例示するように、その主辞－修飾語関係に従って２分木で表現することができる。本研究では、日本語タームJと英語タームEが同型である、すなわち同一の２分木で表現されるときまたそのときに限って、JはEに合成的に翻訳することができると仮定する。この仮定に基づいて、日本語のタームすなわち単語列Jから英語の単語列Eへの合成翻訳の確信度スコア$S(J,E)$を次のように定義する。

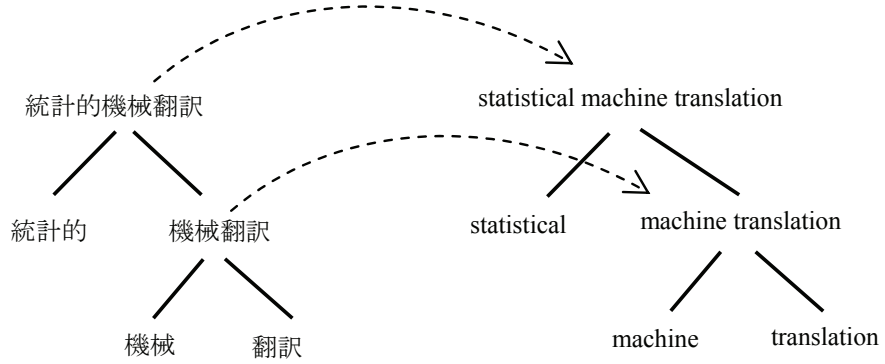
$$S(J,E)=\begin{cases}\lambda\cdot S'(J,E)+(1-\lambda)\cdot C(J,E) & (|J|\geq 2,|E|\geq 2)\\ C(J,E) & (\min\{|J|,|E|\}=1)\end{cases}, \quad [2]$$

ここに、$S'(J,E)$は合成翻訳に基づく確信度スコア、$C(J,E)$は対応する文書の組における共起に基づく相関であり、λは$S'(J,E)$と$C(J,E)$の重みを調整するパラメータ、$|J|$と$|E|$はそれぞれ単語列JとEの長さである。

合成翻訳に基づく確信度スコアは次式で定義する、



(a) Example 1



(b) Example 2

図 1 タームの構造と合成翻訳

$$S'(J,E)=\max_{\substack{1\le i<p\\1\le j<q}}\frac{2\cdot S(jw_1^i,ew_1^j)\cdot S(jw_{i+1}^p,ew_{j+1}^q)}{S(jw_1^i,ew_1^j)+S(jw_{i+1}^p,ew_{j+1}^q)},\quad [3]$$

ここに、$J=jw_1jw_2\cdots jw_p(=jw_1^p)$、$E=ew_1ew_2\cdots ew_q(=ew_1^q)$である。この式は次のような考え方に基づいている。合成翻訳に基づく確信度スコアを二つの構成要素の翻訳の確信度スコアの調和平均として定義する。しかしながら、JとEの正しい構造は不明である。そこで、JとEの可能な分割のすべての組合せに対して確信度スコアを計算し、正しい構造の組合せは確信度スコアを最大にするという仮説に基づいて、確信度スコアの最大値を選択する。

式[3]は、日本語タームとその英語訳語の間で語順が一致するという仮定を示している。語順の一致は一般には成立しない。語順の変化を扱うことができるように式[3]を修正することは難しいことではない。また、式[3]は、二つの構成要素の訳語間の結合可能性を表すファクターを含んでいない。構成要素の訳語の結合可能性は相関 $C(J,E)$にある程度反映されていることを付け加えておく。

Input term: $jw_1^N (= jw_1 jw_2 \cdots jw_N)$

Bilingual lexicon: L

Triangular matrix: $A(i,j)$ (i=1,…,N; j=i,…,N)

Algorithm:

1) For i:=1 to N do

For j:=i to N do

$A(i,j) \leftarrow \phi$.

2) For i:=1 to N do

$A(i,i) \leftarrow \{\langle ew, S(jw_i, ew)\rangle \mid (jw_i, ew) \in L\}$

3) For r:=1 to N–1 do

For i:=1 to N–r do

For k:=i to i+r–1 do

$$A(i,j) \leftarrow A(i,j) \cup \left\{ \langle e_1 e_2, S(jw_i^j, e_1 e_2)\rangle \mid \begin{matrix} \langle e_1, S(jw_i^k, e_1)\rangle \in A(i,k), \\ \langle e_2, S(jw_{k+1}^j, e_2)\rangle \in A(k+1, j) \end{matrix} \right\}$$

図2　合成訳語生成アルゴリズム

次に、動的計画法による合成訳語生成アルゴリズムについて述べる。これは、図２に示すように、文脈自由文法に対する CKY パージングアルゴリズムに類似している。すなわち、各セル $A(i,j)$ が入力タームの部分単語列 jw_i^j に対応し、部分単語列に対する訳語候補とその確信度スコアを記憶するような三角行列を対角線側から計算する。組合せ的な爆発を防ぐため、各セルに記憶する訳語候補を確信度スコアが高い N_2 個に制限することとする。（第５節で述べる実験では、N_2 を 100 とした。）

2.3.5　実験

2.3.5.1　実験方法

JST（科学技術振興機構）の日英科学技術文献抄録コーパスを用いた実験を行った。このコーパスはさまざまなコンパラビリティの日英抄録対から構成されている。日本語の抄録が英語に翻訳された論文もあれば、日本語の抄録とは無関係に英語の抄録が作成された論文もある。日本語抄録の長さは 500～1,000 字程度、英語抄録の長さは 100～300 語程度である。実験では、1980－2004 年の情報工学分野の文献抄録 107,979 対を用いて、相関付き対訳辞書を生成した。日本語と英語のテキストを単語に分割するため、日本語形態素解析器 Mecab[1] と言語非依存の品詞タガー TreeTagger[2] をそれぞれ用いた。

1 http://mecab.sourceforge.net/

2 http://www.ims.stuttgart.de/projekte/corplex/TreeTagger/

提案したフレームワークを評価するため二つのテストセットを用意した。一つは、「人工知能学事典」 (人工知能学会 2008) の和英索引から 1,094 の日本語タームとその英語レファレンス訳を集めた AI テストセットである。もう一つは、「言語処理学事典」 (言語処理学会 2010) の和英索引から 1,661 の日本語タームとその英語レファレンス訳を集めた NLP テストセットである。

二つのテストセットの日本語タームに対し、以下の三つの対訳辞書を用いた要素合成法によって英語訳語のランク付きリストを生成し、比較した。

(1) コーパスから生成した辞書＋通常の辞書

JST コーパスから生成した対訳辞書を EDR 日英辞書[3]、EDICT 日英辞書[4]、英辞郎英日辞書[5]とマージした。通常の辞書では対訳に相関値が与えられていないので、すべての対訳に均一な値 0.1 を与え、コーパスから生成した辞書と通常の辞書の両方に含まれる対訳については二つの値のうち大きい値を採用した。

(2) コーパスから生成した辞書

JST コーパスから生成した対訳辞書のみ

(3) 通常の辞書

EDR 日英辞書、EDICT 日英辞書、英辞郎英日辞書を一つにマージした。訳語のランク付きリストを出力することができるように、日本語と英語の語の組の相関として、JST 文献抄録コーパス中でそれらの語が共起する日英抄録対の数に比例する値を与えた。

人工知能学事典の和英索引からテストセットに含まれない日本語タームとその英語訳語を集め、これを用いて三つの対訳辞書それぞれを用いる場合のパラメータλの値を決定した。(1)コーパスから生成した辞書＋通常の辞書、(2)コーパスから生成した辞書、(3)通常の辞書に対するλの値はそれぞれ 0.40、0.43、0.33 となった。

2.3.5.2 実験結果

表 1 に、３とおりの対訳辞書を用いた要素合成法の各々における正しい訳語の MRR (Mean Reciprocal Rank) と Top k の精度（k=1, 3, 10）を示す。MRR は、正しい訳語のランクの平均値である。Top k の精度とは、入力タームのうち正しい訳語が確信度スコア上位 k 位以内に入ったものの割合である。なお、レファレンス訳と一致する訳語のみを正しい訳語と判定した。この結果は提案したフレームワークが有望であることを示している。すなわち、コーパスから生成した辞書＋通常の辞書の場合だけでなくコーパスから生成した辞書のみの場合も、通常の辞書の場合より精度が向上している。表 1 では、正しい訳語を対訳辞書が与えた訳語の場合と合成的に生成された訳語の場合に分類した。コーパスから生成した辞書＋通常の辞書を用いた場合とコーパスから生成した辞書を用いた場合、正しい訳語の約 30%が合成的に生成された訳語であった。このことは on the fly に訳語を合成することの必要性と有効性を表している。

せいぜい 50%という Top k の精度は、パラレルコーパスあるいはコンパラブルコーパスからの対訳辞書獲得に関する論文で報告されている値と比べてたいへん低い。精度が低い理由の一つは

[3] http://www2.nict.go.jp/r/r312/EDR/index.html
[4] http://www.csse.monash.edu.au/~jwb/edict.html
[5] http://www.alc.co.jp/

表１　実験結果のまとめ

(a) Artificial Intelligence domain (# of test terms: 1094)

Bilingual Lexicon	Corpus-derived + ordinary	Corpus-derived	Ordinary
MRR	0.44	0.4	0.22
Top 1 precision	0.402	0.370	0.197
(Bilingual lexicon)	(0.289)	(0.263)	(0.089)
(Compositional translation)	(0.113)	(0.107)	(0.108)
Top 3 precision	0.464	0.428	0.238
(Bilingual lexicon)	(0.326)	(0.297)	(0.112)
(Compositional translation)	(0.138)	(0.131)	(0.125)
Top 10 precision	0.510	0.473	0.351
(Bilingual lexicon)	(0.351)	(0.320)	(0.135)
(Compositional translation)	(0.169)	(0.153)	(0.144)

(b) Natural Language Processing domain (# of test terms: 1661)

Bilingual Lexicon	Corpus-derived + ordinary	Corpus-derived	Ordinary
MRR	0.35	0.31	0.20
Top 1 precision	0.314	0.282	0.167
(Bilingual lexicon)	(0.231)	(0.202)	(0.102)
(Compositional translation)	(0.083)	(0.081)	(0.066)
Top 3 precision	0.377	0.331	0.217
(Bilingual lexicon)	(0.272)	(0.229)	(0.143)
(Compositional translation)	(0.105)	(0.102)	(0.074)
Top 10 precision	0.415	0.362	0.271
(Bilingual lexicon)	(0.296)	(0.246)	(0,178)
(Compositional translation)	(0.120)	(0.117)	(0.093)

対訳辞書の生成に用いたコーパスとは独立に用意されたテストセットにある。実際、AI テストセットの日本語タームの 11%、NLP テストセットの日本語タームの 12%が、コーパスから生成された対訳辞書にカバーされない単語列を含んでいた。そのようなタームの多くは、あまり使用されない翻字語（例：“タクタイルボコーダ”）、固有名詞を含む語（例：“ボールドウィン効果”）、ほとんど使用されない語（例：“ブラーフミ文字”）であった。

表１の値はやや特異である。正しい訳語が Top 10 に入った入力タームの 80%は、実は正しい訳語が第１位であった。提案方法はコーパス中にあまり出現しないタームに対して信頼できないが、ある程度頻繁に出現するタームに対しては信頼できるといえる。NLP テストセットに対する結果は AI テストセットに対する結果よりかなり悪かった。おそらく、JST コーパスには自然言語処理に関する論文の抄録が比較的少なかったためであろう。

いくつかの入力タームに対し、コーパスから生成した辞書＋通常の辞書を用いた翻訳結果と通常の辞書を用いた翻訳結果を表２に示す。これらの例は提案方法の有効性を示すとともに改良の

表２　要素合成法による翻訳結果の例

#	Input term	Rank	Corpus-derived + ordinary		Ordinary	Reference translation
			Translation	Score	Translation	
1	属性継承 <*ZOKUSEI KEISHOU*>	1	attribute inheritance	0.060	attribute inheritance	property inheritance
		2	attribute succession	0.023	***property inheritance***	
		3	decision tree inheritance	0.021	characteristic inheritance	
2	単純再帰ネットワーク <*TANJUN SAIKI NETTOWAKU*>	1	simple recursive network	0.021	-	simple recurrent network
		2	simple recursion network	0.018	-	
		3	simple recursive service	0.017	-	
3	統合データベース <*TOUGOU DETABESU*>	1	***integrated database***	0.188	integration data base	integrated database
		2	intermolecular	0.069	synthesis data base	
		3	information database	0.058	fusion data base	
4	統計的機械翻訳 <*TOUKEI TEKI KIKAI HONYAKU*>	1	***statistical machine translation***	0.062	statistic object machine translation	statistical machine translation
		2	statistical method machine translation	0.047	statistic target machine translation	
		3	statistical machine translation system	0.046	statistic aim machine translation	
5	統計的統語解析 <*TOUKEI TEKI TOUGO KAISEKI*>	1	statistical syntactic analysis	0.040	-	statistical parsing
		2	statistical method syntactic analysis	0.033	-	
		3	statistical syntactic structure	0.032	-	
6	反駁 <*HANBAKU*>	1	PAC learning model	0.089	counterblast	refutation
		2	・F・	0.067	negation	
		3	***refutation***	0.062	rebuttal	
7	ベイズ決定理論 <*BEIZU KETTEI RIRON*>	1	***Bayes decision theory***	0.056	-	Bayes decision theory
		2	unknown datum theory	0.034	-	
		3	Bayesian decision theory	0.034	-	
8	命題様相論理 <*MEIDAI YOUSOU ROMMRI*>	1	proposition modal logic	0.062	proposition aspect logic	propositional modal logic
		2	***propositional modal logic***	0.036	problem aspect logic	
		3	proposition modal	0.032	proposition state logic	

[Note] Bold and Italicized translations were judged as correct.

余地を示している。

2.3.6　議論

要素合成法は、複合語に限定されるが、コーパスからの語の対訳抽出に広く利用されてきた。通常は既存の対訳辞書を参照して訳語候補を生成し、コーパスを用いて検証する方法をとる (Cao and Li 2002; Tanaka 2002; Baldwin and Tanaka 2004; Tonoike et al. 2006)。これに対し、本稿ではコーパスから生成した対訳辞書を参照することを提案した。実験の結果、このフレームワークによっ

て正しい訳語が生成される可能性が高まることを実証した。正しい訳語が生成されなければ検証手続きは無意味であることに注意すべきである。本稿で提案した改良された要素合成法の大きな特徴は訳語候補の確信度スコアを求めることである。要素合成法においてスコアを用いた研究は既にあるが (Tonoike et al. 2006)、我々のスコアはコンパラブルコーパスに基づくものであるという点でユニークである。

本稿で提案したフレームワークの改良方向について以下に述べる。

第一に、相関付き対訳辞書を改良することが必要である。現在のコーパスから獲得した対訳辞書は非常に多くの誤った対訳を含んでいる、表２の例によると、(属性，decision tree)、(ネットワーク，service)、(反駁，PAC learning model）というような対訳が含まれている。これは、種辞書を使用しないのでやむを得ないことではある。しかし、いったん対訳辞書が獲得されたら、それを使ってよりノイズの少ない対訳辞書を獲得することができる。言い換えると、対訳辞書を逐次的に洗練していくことが可能である。

第二に、確信度スコアを洗練する余地がある。現在のところ、構成要素の訳語の間の関係すなわち結合可能性を考慮していない。確信度スコアの改良の一つの可能性として、構成要素の訳語の確信度スコアの調和平均と構成要素の訳語の間の相関の積を求めることが考えられる。ここで、構成要素の訳語の間の相関はターゲット言語の単言語コーパスから推定することができる。この改良には代替案がある。すなわち、ありそうもない訳語も候補として生成し、ターゲット言語の単言語コーパスあるいは Web を用いて検証する方法でもよいかもしれない (Dagan and Itai 1994)。

第三に、語順の変化を許すように合成翻訳モデルを拡張することが必要である。例えば、日本語タームは名詞列であるがその英語訳語が前置詞句を含むことがある。確信度スコアに構造変換のファクターを組み入れるべきである。いくつかの先行研究において語順の変化を伴う要素合成法が検討されている (Baldwin and Tanaka 2004)。

2.3.7 おわりに

コンパラブルコーパスを用いて要素合成法によるターム翻訳を改良した。対応づけられた文書対からなる２言語コーパスから、ターム中の単語列の対訳とその相関値からなる対訳辞書を獲得する。２言語の単語列の間の相関は、文書対中の共起に基づいて計算される。そして、入力タームに対して、構成要素の間の相関に基づいて定義される確信度スコアとともに訳語候補を合成的に生成する。このようにして、入力タームに対するできるだけ多くの訳語候補の中から正しい訳語を選択することができる。

日英の科学技術文献抄録からなるコンパラブルコーパスを用いた実験を行い、コーパスから獲得した対訳辞書を用いた要素合成法は通常の対訳辞書を用いた要素合成法より高い性能をもつことを実証した。今後の課題として、相関付き対訳辞書を逐次的に改良する方法、確信度スコアの精密化、語順の変化を許す合成翻訳モデルの拡張があげられる。

謝辞： JST 日英科学技術文献抄録コーパスの研究利用をご許可いただいた科学技術振興機構に感謝致します。なお、本研究の一部は科研費（22320032）の助成を受けて実施した。

参考文献

Andrade, Daniel, Tetsuya Nasukawa, and Jun'ichi Tsujii. 2010. Robust measurement and comparison of context similarity for finding translation pairs. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 19-27.

Baldwin, Timothy and Takaaki Tanaka. 2004. Translation by machine of complex nominals: Getting it right. In *Proceedings of the 2nd ACL Workshop on Multiword Expressions: Integrated Processing*, pages 24-31.

Cao, Yunbo and Hang Li. 2002. Base noun translation using Web data and the EM algorithm. In *Proceedings of the 19h International Conference on Computational Linguistics*, pp. 127-133.

Dagan, Ido, and Alon Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, vol. 20, No. 4, pp. 563-596.

Fung, Pascale and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pp. 414-420.

Ismail, Azniah and Suresh Manandhar. 2010. Bilingual lexicon extraction from comparable corpora using in-domain terms. In *Proceedings of the 23rd International Conference on Computational Linguistics*, Poster Volume, pages 481-489.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the ACL*, pp. 48-54.

Matsumoto, Yuji, and Takehito Utsuro. 2000. Lexical knowledge acquisition. In *R. Dale, H. Moisl, and H. L. Somers (ed.). Handbook of Natural Language Processing*, Ch. 24, pp. 563-610 (Marcel Dekker Inc.).

Morin, Emmanuel and Emmanuel Prochasson. 2011. Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora, ACL 2011*, pages 27-34.

Och, Franz Josef, and Hermann Ney. 2003. A Systematic comparison of various statistical alignment models. *Computational Linguistics*, vol. 29, No. 1, pp. 19-51.

Rapp, Reinhard. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 320-322.

Tanaka, Takaaki. 2002. Measuring the similarity between compound nouns in different languages using non-parallel corpora. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 981-987.

Tonoike, Masatsugu, Mitsuhiro Kida, Toshihiro Takagi, Yasuhiro Sasaki, Takehito Utsuro, Satoshi Sato. 2006. Comparative Study on Compositional Translation Estimation using a Domain/Topic-Specific Corpus collected from the Web. In *Proceedings of the 2nd International Workshop on Web as Corpus*, pp. 11-18.

Utsuro, Takehito, Takashi Horiuchi, Kohei Hino, Takeshi Hamamoto, and Takeaki Nakayama. 2003. Effect of cross-language IR in bilingual lexicon acquisition from comparable corpora. In *Proceedings of the 10th Conference of the European Chapter of the ACL*, pp. 355-362.

言語処理学会(編). 2010. 言語処理学事典. 共立出版.

人工知能学会(編). 2008. 人工知能学事典. 共立出版.

3. Automatic Acquisition of Bilingual Technical Terminology Pairs

Kyoto University Denny Cahyadi
Toshiaki Nakazawa
Sadao Kurohashi

3.1 Introduction

Collecting a large number of technical terms is a challenging task. Compared to non-technical terms, the number of technical terms appearing in general documents is relatively small. Technical terms mostly appear in technical documents, such as research papers and patent documents. To collect a large number of technical terms we need a large number of such documents. Unfortunately, those documents are usually not available in full for free. Thus, we are looking for the possibilities to collect technical terms from the free part of technical documents. We found that while most of research papers documents requires subscription fee to be accessed, the abstracts of research papers are usually available for free. The abstract part of a research papers usually contains some essential technical terms written as keywords. Since these keywords are written by the expert (author of the research paper), we assume that these are high quality technical terms. We plan to collect technical terms from this part.

We also found that keywords in some research papers are written in two different languages (usually the original language and English). This opens the possibility to collect non-English-English bilingual technical term pairs. Moreover, since some keywords have common English translations, it is also possible to collect non-English-non-English technical term pairs. It could be done through pivoting, by using English as the pivot language. In our experiment, we tried to collect Chinese-Japanese technical term pairs from abstracts of research papers in Chinese and Japanese language. We first collected Chinese-English and Japanese-English pairs and finally align them to obtain Chinese-Japanese pairs.

3.2 Workflow

To collect Chinese-Japanese technical term pairs, we conduct some experiments with work flow shown by Figure 1. We divide our experiments into three main parts: extracting abstracts and keywords, aligning keywords within a document, and aligning keywords across documents. The details of each part are explained in the following sections:

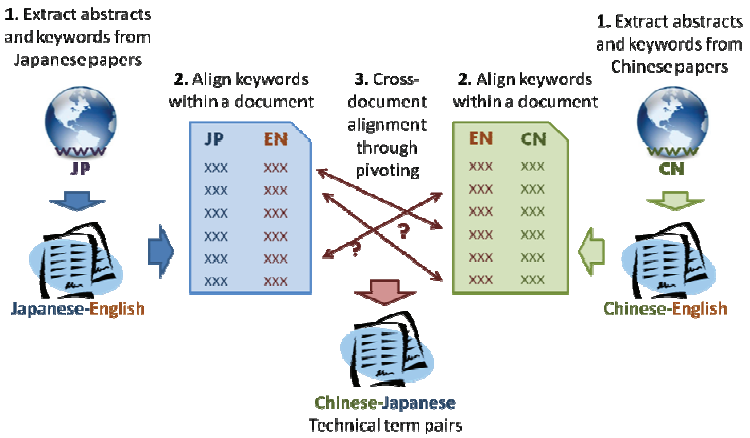


Figure 1: Proposed work flow

3.2.1 Extracting Abstracts and Keywords

The first step of our experiment is collecting a large number of abstracts of research documents. To obtain such documents written in Chinese, we crawl around 170k documents from Chinese research portal CNKI[1]. Since the portal provides search feature, we can easily get the link of each individual document from the search index page. We found that not all documents contain Chinese and English keyword pairs. The documents which have English-only title are very likely to contain English-only keywords. Therefore, we only crawl the documents which contain at least one Chinese word in the title. Out of 170k documents, only 75k documents contain keywords both in Chinese and English. After crawling process completed, the keyword list part of the abstract is extracted. The position of keyword list can be determined by locating the HTML tag which corresponds to the keyword list.

For Japanese side, we use Japanese research paper dataset provided by NII[2]. In contains the XML version of the research paper data available at Japanese research portal CiNii[3]. Since the data is already tagged, we can easily locate the keyword list part by examining the XML tag. This dataset contains about 4.2M of research papers, however only 750k of them contains keywords in both Japanese and English. For the further experiment, only these 750k of Japanese documents and 75k of Chinese documents are used.

3.2.2 Keywords alignment within a document

The second step of our experiment is aligning original-language keywords with English keywords within each document. We consider several methods for the alignment task: 1) monotonic, 2) using SMT tool GIZA++ [1], 3) alignment based on log-likelihood score, and 4) multi-tier alignment.

Monotonic alignment was originally proposed by Ren et al [3]. This alignment method is done based on assumption that keywords in English are likely to be written in the same order with the keywords in the original language. The method is very simple: alignment is done according to the position of the keyword in the keyword list (i.e. the first appearing original language keyword is aligned to the first appearing Chinese keyword and so on).

We are unsure whether this assumption is true for our dataset. We think that some documents may contain keywords written in different order for each language. Therefore, we consider a method which allows reordering during the alignment process. In the second experiment, we use GIZA++ for the alignment. GIZA++ is an alignment tool commonly used for SMT. Given a set of parallel sentence, GIZA++ can create translation table and select the best alignment for each word. To apply GIZA++ in our case, we treat the list of keywords as a sentence and the whole keywords as a single word (by replacing space with underscore), and have GIZA++ to do the alignment.

GIZA++ is designed to align words between languages which have certain rules (grammars). We think that our case may be different. There is no certain rule of how keywords are written. The author of the paper can write the keywords as he/she wish, thus the order of keywords may be random. In order to handle

[1] http://www.cnki.net/
[2] http://www.nii.ac.jp/
[3] http://ci.nii.ac.jp/

this case, we conduct the third experiment. In this experiment, we aligned keyword based on their likelihood score. This is based on an assumption that keyword pairs which often occurring together is likely to be translation of each other. We computed the likelihood score based on a formula introduced by Rapp [2], as seen in Figure 2.

Based on our observation after keyword extraction, we found that the variation of keywords is high, but the occurrence frequency is low. Statistical method such as log-likelihood score may not be able to align keywords correctly due to data sparseness. However, we also found that most keywords consist of more than one word. The variation of these single words is lower and their frequencies are higher than the frequency of the whole keywords. By using statistical information not only from the whole keyword but also from each single word, the sparseness problem may be able to be solved.

$$-2\log\lambda = \sum_{i,j\in\{1,2\}} k_{ij}\log\frac{k_{ij}N}{C_iR_j}$$
$$= k_{11}\log\frac{k_{11}N}{C_1R_1} + k_{12}\log\frac{k_{12}N}{C_1R_2} +$$
$$k_{21}\log\frac{k_{21}N}{C_2R_1} + k_{22}\log\frac{k_{22}N}{C_2R_2}$$

where

$C_1 = k_{11} + k_{12}$ $\quad$ $C_2 = k_{21} + k_{22}$

$R_1 = k_{11} + k_{21}$ $\quad$ $R_2 = k_{12} + k_{22}$

$N = k_{11} + k_{12} + k_{21} + k_{22}$

k_{11} = freq. of common cooccurrence of keyword 1 and keyword 2

k_{12} = freq. of keyword 1 – k_{11}

k_{21} = freq. of keyword 2 – k_{11}

k_{21} = freq. of all word – freq. of keyword 1 – freq. of keyword 2

Figure 2. Log-likelihood formula

In the fourth experiment, we use statistical data from every single word instead of the whole keyword. We use multi-tier alignment method to align the keywords. Figure 3 illustrates how this method works. First, each keyword in the original language is considered as a possible translation candidate of each English keyword. Every possible combination is then generated. For each combination, every keyword is segmented into several single words. Next, all possible pairing for every segment is generated. For example, for *Harvesting robot* keyword from *combination #2*, there are two possible pairing, (収穫-*Harvesting*, ロボット-*robot*) and (収穫-*robot*, ロボット-*Harvesting*). At this level, statistical information of each single word is used to determine whether the pairing is a correct translation or not. A pair which is likely to be a translation of each other has a higher score than a pair which is unlikely to be a translation of each other, for example (収穫-*Harvesting*, ロボッ-*robot*) has a higher score than (収穫-*robot*, ロボッ-ト *Harvesting*). The pairing which has the highest score is selected. Score from this pairing is then used as the score of the corresponding keyword. After the score of every keyword is computed, the score of a combination is computed as the summation of the score of all keyword it contains (e.g. the score of *combination #2* is the summation of the score of [収穫ロボット-*Harvesting robot*] and [超音波センサー-*ultrasonic sensor*]). Finally, a combination with the highest score is selected as the best translation candidate. From this combination, alignment for each keyword can be determined.

As mentioned above, statistical information of every single word is used to determine the pairing. For this purpose, we run GIZA++ again, but this time with all keyword segmented into single word. GIZA++ then computes the translation probability of each word and we use this value to determine the pairing.

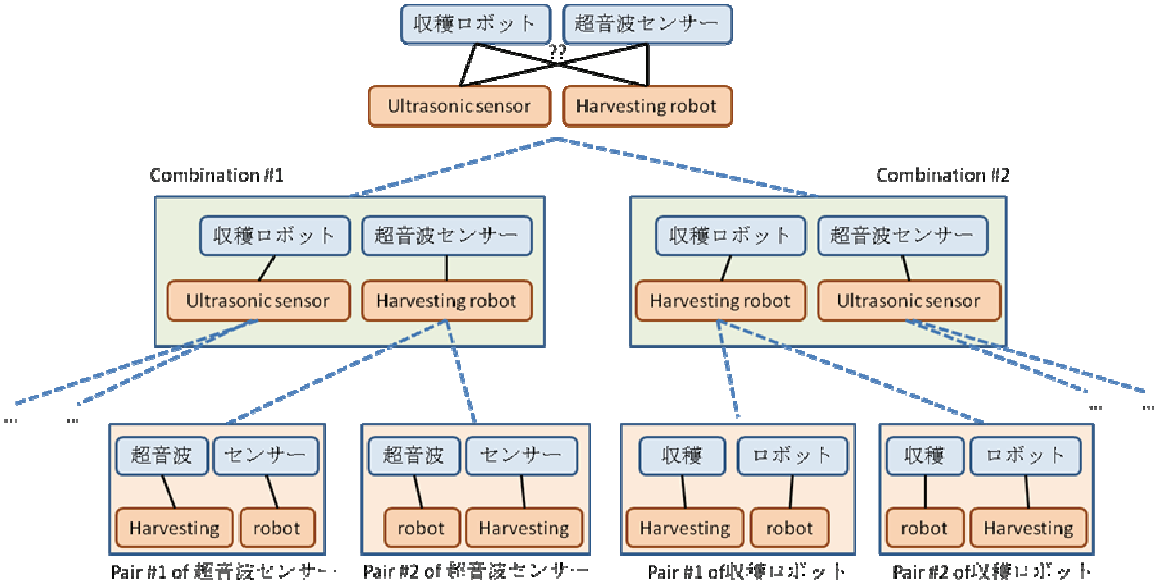


Figure 3. Illustration of multi-tier alignment

3.2.3 Keywords alignment between Chinese and Japanese

After all keywords within every document are aligned, we obtain Chinese-English and Japanese-English keyword pairs. For the next step, we align these keywords to obtain Chinese-Japanese pairs. Our method is very simple; we just align keywords which have a similar English translation. We found there were minor variations in English keywords with the same meaning (e.g. *broad-band noise* and *broadband noise*). If we use an exact match to align, we cannot tolerate any variations (which maybe correct) and may affect the final alignment. Therefore, normalized edit distance score is used instead of exact matching. Normalized edit distance is defined as the number of operation required to convert a string into another (to make them similar) divided by its length. Small variations of English keywords are allowed and treated as the same keywords if their normalized edit distance is lower than a threshold.

3.3 Experiment and result

For all experiments, 752,945 Japanese documents and 75,398 Chinese documents which contain bilingual keywords are used. On average there are 2 to 10 keywords in each document. For evaluation, bilingual keyword pairs obtained by our methods are compared to manually compiled technical terms dictionaries. The keywords are labeled as *correct* if they are found on dictionary and its translation is the same to the entry on the dictionary, as *incorrect* if they are found but the translation are differ, and as *not found* if they are not found on the dictionary.

Four different methods are used in our within document alignment experiments: monotonic alignment, alignment using GIZA++, log-likelihood score-based alignment, and multi-tier alignment. The comparison of *correct* result for each method is shown in Table 1. The results only differ slightly for each method. Monotonic alignment produces the best result for Japanese-English alignment and alignment by GIZA++

produces the best result for Chinese-English alignment. Log-likelihood score-based alignment produces the worst result for Chinese-English alignment and multi-tier alignment produce the worst result for Japanese-English alignment.

After keyword alignment for each document is completed, we aligned Chinese and Japanese keyword via English. The result of the alignment is shown in Table 2. In Table 2, the proportion of *correct*, *incorrect*, and *not found* for both within document alignment and Chinese-Japanese alignment is shown. Within document alignment result is taken from the result of monotonic alignment experiment.

Alignment Method	**JP-EN**	**CN-EN**
Monotonic	**59,630**	12,044
GIZA++	59,258	**12,486**
Log-Likelihood	56,453	3,001
Multi-tier	43,123	10,923

Table 1. Number of correct alignment for within-document alignment

	JP-EN	**CN-EN**	**CN-JP**
Correct	59,630	12,044	4,695
Incorrect	20,878	5,228	15,955
Not found	227,834	80,059	19,049
Total	308,342	97,331	39,699

Table 2. Proportion of correct, incorrect and not found

Compared to the total number of keyword pairs, the number of correct pairs is relatively small. For Japanese-English, it is only about 19%, for Chinese-English it is about 12% and for Chinese-Japanese is only about 11%. We think that our method so far is not very effective for the keyword alignment task with sparse data. However we also consider that the manually compiled dictionary contains fewer entries than our total keywords or has different coverage. As seen in Table 2, the number of *not found* keywords is far larger than the number of *correct*/*incorrect* keywords. We think that some keywords with *not found* category may actually correct keywords. Table 3 shows some example of keywords we get from our experiments.

Chinese	**English**	**Japanese**	**Result**
口蹄疫病毒	foot-and-mouth disease virus	口蹄疫ウイルス	correct
紫外吸收光谱	ultraviolet absorption spectrum	紫外吸収スペクトル	correct
差向异构化	epimerization	エピマー化	correct
肠肌丛	myenteric plexus	筋層間神経叢	correct
电导滴定	conductometric titration	導電率滴定	correct
草莓	strawberry	収穫ロボット	incorrect
振动控制	vibration control	可変減衰器	incorrect
近红外光谱	near infrared reflectance spectroscopy	黒ボク土壌	incorrect
挥发性有机化合物	volatile organic compounds	性有機化合物	not found
双曲型偏微分方程	hyperbolic partial differential	双曲型偏微分方程式	not found
类金属硫蛋白	metallothionein	メタロチオネイン	not found
骨性关节炎	osteoarthritis	変形性関節症	not found

Table 3. Example of *correct*, *incorrect*, and *not found* keywords

As seen in Table 3, some pairs categorized as *not found* is actually *correct* keyword pairs (e.g. 类金属硫蛋白 is a correct translation of メタロチオネイン). We think that the actual number of correct pairs is actually larger than the number shown in Table 2. In the future, we are planning to use different dictionary and human evaluation for better result.

Based on our observation, incorrect Chinese-Japanese pair is often caused by one of incorrect alignment between Chinese-English or Japanese-English. On the 6th line of Table 3, on the Chinese side, 草莓 is correctly aligned to *strawberry*. However, *strawberry* is misaligned to 収穫ロボット(*harvesting robot*) on the Japanese side. As a result, the final alignment becomes incorrect. For further analysis, we examine the documents where these keywords are written. We found that in the Japanese document, there are 収穫ロボット and いちご keywords with their corresponding English keywords *harvesting robot* and *strawberry*. However, the corresponding English keywords are written with different order than Japanese one. Our method failed to reorder the keyword, thus the alignment become incorrect. We are planning to improve our method in the future.

3.4 Conclusion

Our experiment shows the possibilities to collect a large number of technical term pairs from abstracts of research papers. However, our present method is not efficient enough to collect them. Monotonic alignment method sometimes fails to get the correct alignment because some keywords in English are not written in the same order with keywords in the original language. Statistical-based method such as log-likelihood and GIZA++ allows reordering of the keywords. However, they sometimes fail due to data sparseness. In the future we would like to try the other methods such as HMM alignment method. We also plan to improve our evaluation method by using human evaluation in the future.

References

[1] Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Association for Computational Linguistics*, 29(1):19–51.

[2] Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 519–526, Stroudsburg, PA, USA. Association for Computational Linguistics.

[3] Feiliang Ren, Jingbo Zhu, and Huizhen Wang. 2010. Web-based technical term translation pairs mining for patent document translation. In *Natural Language Processing and Knowledge Engineering (NLP-KE), 2010 International Conference*, pages 1–8

４．　規則方式機械翻訳と統計的後編集を組み合わせた

特許文の日英機械翻訳(その 4)

山梨英和大学　江原暉将

4.1　はじめに

これまで、規則方式日英機械翻訳(RBMT)と統計的後編集(SPE)を組み合わせることで翻訳精度の向上を図ってきた[江原、小玉 2005][江原 2006][江原 2008][江原 2010] [江原 2011]。これまでのシステム比較を表 1 に示す。BLEU や NIST の評価値が向上してきている。これまでは、後編集として句レベルの SPE(Phrase-based SPE)を用いてきた。現在は、階層的な SPE(Hierarchical SPE)および 構文レベルの SPE(Syntax-based SPE)が利用可能である[Hoang 2009]。そこで、今回は NTCIR-9 のデータを使ってこれら 3 者を比較する。

4.2　本報告で用いる訓練データと試験データ

本報告で用いるデータは、[Ehara, 2011]と同じものである。つまり、国立情報学研究所から「NTCIR-9 特許翻訳タスク参加者用テストコレクション」として提供された NTCIR-9 の PatentMT task, JE subtask の formal run のためのデータである[Goto, 2011]。本報告で用いる試験データは NTCIR-9 で使用した 2000 文全体ではなく、その中から人手評価が行われた 300 文を用いた。訓練データの元データは、日英特許平行コーパスであり、NTCIR-7 で用いられた約 180 万文対と NTCIR-8 で用いられた約 140 万文対、あわせて約 320 万文対から成る。言語モデル(LM)の訓練データとしては、NTCIR-8 で用いられた訓練データの英語部分を抽出して用いた。よって約 140 万文である。翻訳モデル(TM)の訓練データは、[Ehara, 2011]に示した方法によって元データから 291,475 文対の日英対応データを選択して用いた[1]。ただし、Syntax-based SPE では、利用した構文解析器 Enju [Miyao 2008]で構文解析ができなかった 61 文対を除外したため、翻訳モデルの訓練データは 291,414 文対となった。パラメータ調整のための開発データは提供された 2000 文対の冒頭部分 300 文対を利用した。

4.3　実験結果

表 2 に実験結果を示す。Phrase-based SPE から Hierarchical SPE とすることで BLEU、NIST ともに若干向上しているが、Syntax-based SPE では逆に低下している。

4.4　翻訳結果

テストデータに対する翻訳結果の例を付録に示す。表中、src は日本語原文、ref は基準英語訳文、rbmt は規則方式機械翻訳の出力、P-spe は Phrase-based SPE の出力、H-spe は Hierarchical SPE の出力、S-spe は Syntax-based SPE の出力を示す。各例文に対して考察を加える。

例文 1：「図４に」の係り先が正しくは「示している。」であるが rbmt では構文解析を間違えており「流れる」に係るように解釈している。P-spe、S-spe ともにこの解釈を訂正できていない

[1] 翻訳モデルの訓練データ選択では、試験データ全体つまり 2000 文を用いている。本報告で使用した 300 文に限って選択したわけではない。

表 1 規則方式機械翻訳(RBMT)と統計的後編集(SPE)を組み合わせたシステムの推移 [2]

	[江原、小玉2005]	[江原2006]	[江原2008]	[江原2010]
RBMT部分	市販品A	非市販品	非市販品	市販品B
SPE部分	単語レベル(isi)	単語レベル(isi)	句レベル(Moses)	句レベル(Moses)
TM学習器	Giza-pp	Giza-pp	Giza-pp	Giza-pp
TM訓練データ	特開報／PAJ 9万3千文対	特開報／PAJ 9万3千文対	特開報／PAJ 9万3千文対	NII NTCIR-7 8万2千文対
LM学習器	Srilm	Srilm	Srilm	Srilm
LM訓練データ	PAJ 33万文	PAJ 33万文	PAJ 33万文	US patent 180万文
BLEU	0.1607	0.1728	0.2912	0.2998
NIST	4.7184	4.7893	6.3398	7.3058

表 2 実験結果

	Phrase-based SPE	Hierarchical SPE	Syntax-based SPE
RBMT部分	市販品B	市販品B	市販品B
SPE部分	Moses	Moses	Moses
TM学習器	Giza-pp	Giza-pp	Giza-pp
TM構築の構文解析器	― ― ―	― ― ―	Enju 2.3 (moguraを使用)
TM訓練データ	NII NTCIR-7 and 8 291,475文対	NII NTCIR-7 and 8 291,475文対	NII NTCIR-7 and 8 291,414文対
LM学習器	Srilm	Srilm	Srilm
LM訓練データ	US patent 180万文	US patent 180万文	US patent 180万文
BLEU	0.3086	0.3127	0.2391
NIST	6.8189	6.8347	6.2593

が、H-spe は訂正できている。

[2] 使用ツールの詳細は以下のとおり。
言語モデル学習器：http://www.speech.sri.com/projects/srilm/の srilm.tgz ver.1.5.5
翻訳モデル学習器：http://code.google.com/p/giza-pp/の giza-pp-v1[1].0.1.tar.gz
単語レベルデコーダ：
http://www.isi.edu/publications/licensed-sw/rewrite-decoder/index.html の
isi-rewrite-decoder-r1.0.0a/linux/decoder.linux.public (現在ダウンロードできないようである)
句レベルデコーダ：http://sourceforge.net/svn/?group_id=171520 の moses.2007-05-29.gz
構文レベルデコーダ：http://sourceforge.net/projects/mosesdecoder/files/
構文レベルモデルのための構文解析器：Enju 2.3, http://www-tsujii.is.s.u-tokyo.ac.jp/enju/index.ja.html
BLEU と NIST の計算プログラム：ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v11b.pl
ただし、BLEU 値を文単位で計算するために計算式を若干変更してある[江原 2007]。

例文 2:「好適」が ref では” appropriate for”と訳されている。rbmt では” preferred as”と訳し、P-spe、H-spe、S-spe ともに” preferable as”と訳している。また、「操舵角速度ｄδｓｗ」は「演算横加速度の時間変化ｄＧｙ２」と並列されるが、rbmt で「操舵角δｓｗ」と並列するように誤って解釈され、全ての spe で訂正できていない。

例文 3:「水素結合性化合物を塗布液に含有せしめ」の部分が rbmt では” Coating liquid is made to contain the hydrogen bond nature compound”となっており、P-spe では” is made to contain”の部分が” is added”となり、H-spe では” is contained in”となり、S-spe では” is made to contain”と rbmt と同一である。

例文 4:「鉛筆を削る」が rbmt では” pencil can be shaved”となり、P-spe では” light beam can be scraped”となり、H-spe では” light beam can be removed”と両者とも改悪されている。S-spe では” pencil can be sharpened”と改善されている。

例文 5:「ワークに加工等の作業を行う」が rbmt、P-spe、H-spe、S-spe ではそれぞれ”working processing etc. to a work”、”operating process or the like to the operation”、”machining process or the like to the operation”、”working process or the like to a work”と訳されている。いずれも訳語が不適切である。ref は”machining or otherwise working a workpiece”である。

例文 6:「半田付ロボット」が rbmt では”robot with Handa”と訳され、P-spe、H-spe、S-spe では、”robot with solder”と訳されている。ref は”soldering robot”である。また、「こて先」が rbmt と H-pse では ”こて point”と日本語が混じり、P-spe と S-spe では”spatulate point”と訳されている。ref は”iron tip”である。

4.5 おわりに

規則方式機械翻訳システム(RBMT)と統計的後編集システム(SPE)を組み合わせて、特許文書用機械翻訳システムを構築している。今回の報告では、SPE 部として従来用いてきた句レベルの SPE (Phrased-based SPE; P-spe)に加えて階層的な SPE (Hierarchical SPE; H-spe)と構文レベルの SPE (Syntax-based SPE; S-spe)の 3 種のシステムを用いた場合の比較を行った。BLUE および NIST の値は、P-spe より H-spe は若干向上し、S-spe では逆に低下した。また翻訳結果のいくつかについて分析を加えた。ただし、詳しい分析は今後の課題である。

参考文献

[越前谷 2009] 越前谷博ほか：NTCIR-7 データを用いた機械翻訳評価規準のメタ評価、平成 20 年度 AAMT/Japio 特許翻訳研究会報告書、pp.2-13, March, 2009.

[江原、小玉 2005] 江原暉将、小玉修司：特許文の日英機械翻訳結果と PAJ を比較して翻訳知識を抽出する研究、平成 16 年度 AAMT/Japio 特許翻訳研究会報告書、pp.86-96, March, 2005.

[江原 2006] 江原暉将：規則方式機械翻訳と統計的後編集を組み合わせた特許文の日英機械翻訳、平成 17 年度 AAMT/Japio 特許翻訳研究会報告書、pp.40-44, March, 2006.

[江原 2007] 江原暉将：新しい機械翻訳自動評価基準を目指して、平成 18 年度 AAMT/Japio 特許翻訳研究会報告書、pp.2-11, March, 2007.

[江原 2008] 江原暉将：句レベルの統計的後編集と翻訳精度の評価、平成 19 年度 AAMT/Japio 特許翻訳研究会報告書、pp.2-11, March, 2008.

[江原 2010] 江原暉将：規則方式機械翻訳と統計的後編集を組み合わせた特許文の日英機械翻訳(その 2)、平成 21 年度 AAMT/Japio 特許翻訳研究会報告書、pp.56-60, March, 2010.

[江原 2011] 江原暉将：規則方式機械翻訳と統計的後編集を組み合わせた特許文の日英機械翻訳(その 3)、平成 22 年度 AAMT/Japio 特許翻訳研究会報告書、pp.43-46, March, 2011.

[Ehara, 2011] Terumasa EHARA : Machine translation system for patent documents combining rule-based translation and statistical post-editing applied to the PatentMT Task, Proceedings of NTCIR-9 Workshop Meeting, pp.623-628, Dec. 2011.

[Goto, 2011] Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita and Benjamin K. Tsou : Overview of the Patent Translation Task at the NTCIR-9 Workshop, Proceedings of NTCIR-9 Workshop Meeting, pp.559-578, December, 2011.

[Hoang 2009] H. Hoang, P. Koehn, and A. Lopez : A Unified Framework for Phrase-Based, Hierarchical, and Syntax-Based Statistical Machine Translation, Proceedings of the International Workshop on Spoken Language Translation, pp. 152-159, 2009.

[Miyao 2008] Yusuke Miyao and Jun'ichi Tsujii : Feature Forest Models for Probabilistic HPSG Parsing, Computational Linguistics, Vol. 34, No. 1, pp. 35-80, MIT Press.

付録 翻訳例

1	src	**図４に**非磁性層を**流れる**電流に対する電気抵抗を縦軸にとってグラフに示している。
	ref	**FIG. 4** is a graph showing an electric resistance to an electric current **flowing** in the nonmagnetic layer with the axis of ordinates representing the electric resistance.
	rbmt	The electrical resistance to the current **which flows into Drawing 4** through a non-magnetic layer is shown in the graph for the vertical axis .
	P-spe	The electric resistance to the current **flowing in the FIG. 4** through the nonmagnetic layer is shown in the graph on the vertical axis.
	H-spe	The electric resistance to the current **flowing** through the nonmagnetic layer is shown in the graph, the vertical axis **in FIG. 4**.
	S-spe	The electrical resistance to the current **which flows to FIG. 4** through a nonmagnetic layer is shown in the graph, the vertical axis.

2	src	つまり、**操舵角速度ｄδｓｗ**、又は操舵角δｓｗから計算される演算横加速度の時間変化ｄＧｙ２が、ロール入力速度ｄＲｍとして**好適**である。
	ref	That is, the **steering angular velocity (d sw)**, or variation in time (dGy2) of the estimated lateral acceleration calculated from the steering angle (δsw) is **appropriate for** the roll input velocity (dRm).
	rbmt	That is , time change dGy2 of the operation lateral acceleration calculated from **steering angular velocity ddeltasw** or steering angle deltasw is **preferred as** roll input speed dRm .
	P-spe	That is, the change dGy2 of the operation lateral acceleration calculated from the **steering angular velocity ddeltasw** or steering angle deltasw is **preferable as** roll input speed dRm.
	H-spe	That is, the time change dGy2 of the lateral acceleration calculated from the **steering angular velocity ddeltasw** and the steering angle deltasw is **preferable as** the roll input speed dRm.
	S-spe	That is, the time change dGy2 of the operation lateral acceleration calculated from **steering angular velocity ddeltasw** or steering angle deltasw is **preferable as** roll input speed dRm.

3	src	本発明における**水素結合性化合物**は、還元剤と同様に溶液形態、乳化分散形態、固体分散微粒子分散物形態で**塗布液**に**含有せしめ**、感光材料中で使用することができる。
	ref	The **hydrogen bonding compound** of the invention can be used in the photothermographic material by **being incorporated into** a **coating solution** in the form of solution, emulsion dispersion, or solid fine particle dispersion, similar to the case of the reducing agent.
	rbmt	**Coating liquid is made to contain** the **hydrogen bond nature compound** in the present invention like a reducing agent with a solution form , emulsification distributed voice , and a solid particulate dispersion distribution thing form , and it can be used in photosensitive materials
	P-spe	**The coating liquid is added hydrogen bond-forming compound** of the present invention as a reducing agent in the form of a solution, an emulsion dispersion or a solid dispersed fine particle dispersion and can be used in the photosensitive material.
	H-spe	**The coating liquid is contained** in the **hydrogen bond-forming compound** of the present invention in the form of a solution, an emulsion dispersion, the reducing agent and the solid fine particle dispersion, and can be used in the photosensitive material.
	S-spe	**The coating liquid is made to contain the hydrogen bond-forming compound** in the present invention as a reducing agent with a solution form, emulsified dispersion voice, and a solid microparticle dispersion, and it can be used in the photosensitive material.

4	src	突起を切削機構に深く挿入することにより、鉛筆は切削後すぐに突起に当接するので、芯の太い**鉛筆を削る**ことができる。
	ref	Moving the projection deep into the cutting mechanism enables a pencil to strike on the projection immediately after shaving, thereby enabling **sharpening of a pencil** having a thick core.
	rbmt	Since a pencil contacts a projection immediately after cutting by inserting a projection in a machining device deeply , a **pencil** with a thick core can **be shaved** .
	P-spe	Since the pencil comes into contact with the projection immediately after the cutting by inserting a projection in a machining apparatus and a **light beam** having a thick core can **be scraped** off.
	H-spe	Since the pencil a machining apparatus is in contact with the projection immediately after the cutting by inserting the projection **light beam** by a thick core can **be removed**.
	S-spe	Since a pencil contacts the projection immediately after cutting by inserting a projection in a machining apparatus deeply, a **pencil** with a thick core can **be sharpened**.

5	src	一方、**ワーク**に加工等の**作業を行う**場合、ワークに対して座標系が設定され、このワーク座標系上の位置に基づいて加工がなされる。
	ref	On the other hand, when machining or otherwise **working a workpiece**, a coordinate system is set for the workpiece, and machining or the like is carried out on the basis of positions in the workpiece coordinate system.
	rbmt	On the other hand , when **working processing** etc. to a **work** , a coordinate system is set up to a work and processing is made based on the position on this workpiece coordinate system .
	P-spe	On the other hand, when the **operating process** or the like to the **operation**, a coordinate system is set to the workpiece and processing is carried out based on the position on the work coordinate system.
	H-spe	On the other hand, when the **machining process** or the like to the **operation**, a coordinate system is set to the workpiece and the processing based on the position on the work coordinate system.
	S-spe	On the other hand, when **working process** or the like to a **work**, a coordinate system is set to the workpiece and processing is made based on the position on the workpiece coordinate system.

6	src	作製した半田合金をやに入り半田（やに入り半田のフラックスはJIS Z 3283：2001によるA級品、フラックス含有量3．5%、線径0．8mm）に加工し、**半田付ロボット**を用いて、**こて先**にやに入り半田を供給した。
	ref	Each prepared soldering alloy was worked into a resin-containing solder (resin-containing solder flux was an A-class product by JIS Z3283:2001, flux content of 3.5%, linear diameter of 0.8 mm), and the resin-containing solder was supplied to an **iron tip** using a **soldering robot**.
	rbmt	The produced solder alloy was processed into the solder containing resin (Class A article according [the flux of the solder containing resin] to JIS Z 3283 : 2001 , 3.5 % of flux content , 0.8 mm of 線径) , and the solder containing resin was supplied to the **こて point** using the **robot with Handa** .
	P-spe	The prepared solder alloy was processed into a solder containing resin (the class A component by the flux of the solder containing resin) in JIS Z 3283: 2001, 3.5% of the flux amount, 0.8 mm in diameter) and a solder containing resin is supplied to the **spatulate point** using the **robot with solder**.
	H-spe	The prepared solder alloy is processed into a solder containing resin (class A parts of the flux of the solder containing resin in JIS Z 3283: 2001, a solder containing 3.5% of flux amount, 0.8 mm in diameter), and the resin is supplied to the **こて point** using the **robot with solder**.
	S-spe	The produced solder alloy was processed into a solder containing resin & class A article according the flux of the solder containing resin] to JIS Z 3283: 2001, 3.5% of flux content, 0.8 mm of wire diameter ″, and the solder containing resin was supplied to the **spatulate point** using the **robot with solder**.

５．　語のグループ化を用いた特許文動詞の訳し分け

山形大学　横山 晶一
高野　雄一

5.1　はじめに

近年，特許のような知的財産が，社会における貴重な存在として認識されており，これに伴う特許申請数の増加が著しい．また，国際的な特許の共有化に伴い国際特許も増加中にあり，正確で迅速な機械翻訳が求められている．

日英機械翻訳における訳文品質の分析[1]において，訳文品質低下の原因は訳し分けの不適切さであると報告されている．訳し分けとは，ある文中の単語を翻訳するときに訳の候補が複数ある場合，その文に最も適した訳を選択するということである．例えば，「含む」という動詞は，「全体の一部として含む」意味合いの文では”include”，「要素・成分として含む」という意味合いでの文では”contain”と訳される．この訳し分けの精度を向上させるためには，文中で使用された単語の意味（語義）を解析する必要がある．

本研究では単語の意味解釈をした上での訳し分けのために，文章を意味のつながりで示すことの可能な語のグループ化を行う[2]．語のグループ化とは“「男」「少女」を＜人＞と分類，「荷物」「鞄」を＜具体物＞と分類にする”などと，語を分類付ける方法のことをここでは言う．語のグループ化が訳し分けに役立つかどうかを調べ，従来よりも精度の高い訳し分けが可能なシステムを作成する．

本稿は，主として[3]に基づき，その後の成果を[4]に基づいて加筆したものである．

5.2　提案手法

適切な訳し分けを行うためには，文の意味を考慮する必要がある．しかし，詳細説明や要件が長大で難解であるという特許文の特徴から，精度の高い機械翻訳が困難であるという現状がある．本研究では動詞の訳し分けを改善することによる翻訳精度の向上を目標とする．入力テキストから冗長な部分を排除し，動詞の前後のテキストから適切な訳し分けとなる対訳動詞を抽出し，修正する．方法を図 1 に示す．入力テキストから，対象の動詞と動詞前後のテキストを抜き出す．抜き出したテキストをグループ辞書により置換する．置換したテキストを訳し分け辞書のスコアに従って，訳し分け候補動詞のスコアを計算し，スコアが最も高くなった動詞を対訳として出力する．

(1) グループ辞書

グループ辞書とは同一の意味・概念となる語を一つにまとめた辞書である．例えば，「男性」，「女性」，「子供」などの語を＜人間＞というグループにまとめる．語を一つにすることで，意味的に同様である文を同様のものとして扱うことが可能となる．

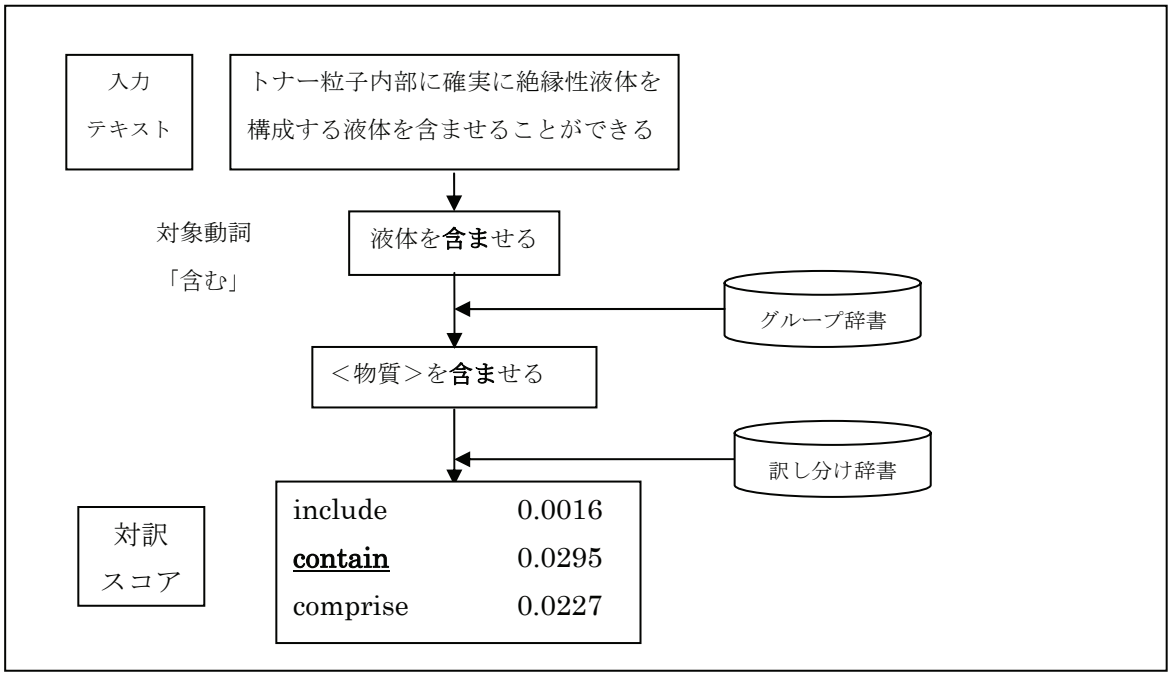


図 1　対訳動詞算出

表 1　動詞訳し分け辞書

動詞	出現数	スコア	テキスト
含む	1071	0.00305	＜道具＞に含まれる
(include)	621	0.00177	＜機械＞に含まれる
	602	0.00171	＜情報＞に含まれる
含む	542	0.00299	＜道具＞に含まれる
(contain)	363	0.00201	＜動物＞に含まれる
	332	0.00183	＜食料＞に含まれる

(2) 動詞訳し分け辞書

訳し分け辞書の作成には対訳付き特許テキストデータとグループ辞書を用いる.

テキストデータから対象の動詞を含めた係り受けや N-gram をとる部分を抜き出す. 抜き出した文を形態素解析し，グループ辞書を用いて各語を置換する．置換した文と，対象の動詞，対訳となる動詞を１つにまとめ，訳し分け辞書に追加する．訳し分け辞書では出現数を数えておき，出現数に応じてそのテキスト形でのスコアを決定する．現在，スコアの算出方法としては，非常に単純に，テキスト出現数と，対訳動詞においてのテキスト総数の商をスコアとして扱っている．訳し分け辞書の作成例を表 1 に示す.

(3) 訳し分け評価

作成した訳し分け辞書を用いて，入力されたテキスト中にある動詞の訳し分け判定を行う．判定ではテキストの置換まで，訳し分け辞書作成と同様の処理を行う．置換を行った後，訳し分け辞書でのスコアを用いて，各対訳英語動詞でのスコアを算出する．スコアの合計が最も高くなった動詞を正しい対訳動詞として出力する．

5.3 実験

本研究のシステムを利用し，日本語テキストを Google 機械翻訳(http://translate．google．co．jp/)に通し，翻訳結果の修正をする実験を行った．

(1) 実験設定

実験の流れを図２に示す。今回訳し分け辞書の作成に用いる学習データとしては，特許明細書文アラインメント[5]の日英対訳テキストからランダムに 500 万文を抽出し用いた．この特許文は日本から米国へ出願された対応特許の明細書の文を NICT の Align で対応付けたものであり，日本語文とそれを人手で英訳した文が収録されている．

グループ辞書には日本語語彙大系[6]を使用する．置換の際には語の上位語に変換する． 基本的には第６層目に変換し，置換元の語がそれ以下の層に該当する語の場合はそのままの形で用いる．

訳し分けの対象として扱う動詞には，出現数が多く複数の訳し分けがある動詞として「含む」(include, contain, comprise), 「得る」(attain, obtain, derive),「用いる」(use, utilize, adopt)を扱った． 訳し分け辞書に登録するテキストには連続する名詞の場合最後の名詞のみを残し，動詞を中心とした単語 5gram を抜き出す前処理を行う．学習用の特許文から各日本語動詞をテキスト中に含み，動詞の対訳が上記の語となる文を収集し，それら用いて訳し分け辞書を作成した．作成した訳し分け辞書から（出現数 / 出現総数）でスコアを算出する．学習数は「含む」の場合，include:89542, contain:35492, comprise:3472 であった．

実験の評価に用いるテキストは，特許明細書文学習データとして，用いていないテキストからランダムに抜き出して用いた．Google 機械翻訳の翻訳結果に対し，本システムによる修正の結果，訳し分けがどの程度できているのか精度を調べた．

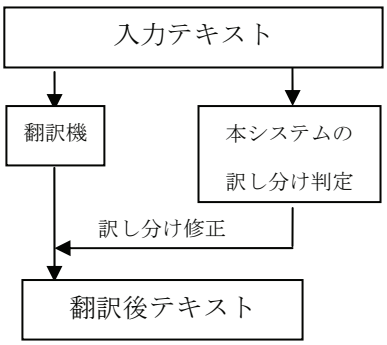


図２　実験の流れ

表２　訳し分け成功例

動詞：正答対訳	含む：contain
入力テキスト	しかし、タッチパネル装置の性質上、画面洗浄剤や汗などのように酸やアルカリ成分を含んだ水分が浸入することが考えられる。
カット処理後	成分を含んだ
訳し分けスコア	###comprise:0.000125271896958171### ###contain:0.000423434622522227### ###include:0.000146979704186892###

表３　訳し分け失敗例

動詞：正答対訳	含む：comprise
入力テキスト	現像手段２は、現像槽２３と、トナープレチャージボックス２２と、トナー補給ボックス２１とを含む。
カット処理後	とを含む。
訳し分けスコア	###comprise:0.00151465111776697### ###contain:0.000167954515078649### ###include:0.000938526513374003###

表4　【含む】の実験結果

	含む		
	include	contain	comprise
正→正	14	23	12
正→誤	9	5	2
誤→正	13	11	22
誤→誤	14	11	14
合計数	50	50	50
正答率	0.5400	0.6800	0.6800

表５　【得る】の実験結果

	得る		
	attain	obtain	derive
正→正	16	22	16
正→誤	6	3	11
誤→正	15	12	7
誤→誤	13	13	16
合計数	50	50	50
正答率	0.6200	0.6800	0.4600

表６　【用いる】の実験結果

	用いる		
	Use	utilize	adopt
正→正	25	0	0
正→誤	13	0	0
誤→正	7	18	22
誤→誤	5	32	28
合計数	50	50	50
正答率	0.6400	0.3600	0.4400

(2) 実験結果

訳し分けの成功例を表２に，失敗例を表３に示す．また，実験の結果を表４～６に示す．「正」は正しい訳，「誤」は誤った訳を表し，「正→誤」は Google 翻訳の出力は正しいが訳し分けの修正の結果判定が誤った，ということを表す．

誤り文の訂正率 50%，正答の修正を含めた全体での正答率は 62%という結果になった．Google 翻訳の訳し分けの正答率が 47%であったことを考慮するとある程度訳し分けを改善できたと考える．

今回は動詞を中心とした 5gram で訳し分け辞書を作成したが，係り受けを考慮することや，スコアの計算方法を見直すこと，特許文に特化したグループ辞書の導入により，訳し分け精度の改善が期待できると考える．

5.4　終わりに

本論文では語のグループ化を用いて特許文動詞の訳し分けをするシステムを作成した．訳し分けの精度を調べるために，Google 機械翻訳の結果を修正する実験を行った．実験の結果，ある程度訳し分けの改善を確認出来た．訳し分け精度の向上のためには特許文用に特化したグループ辞書の作成や，スコア計算方法の改善が必要である．

謝辞

本研究に際し，Japio から，資料の提供を賜りました．ここに感謝の意を表します．

参考文献

[1]麻野間直樹，中岩浩巳：目的言語の単語共起情報を利用した訳語選択と未知語の訳出，言語処理学会第５回年次大会論文集，pp. 442-448，(1999)

[2]S. Yokoyama, Y. Takano: Investigation for Translation Disambiguatioin of Verbs in Patent Sentences using Word Grouping, *Proceedings of the 4th Workshop on Patent Translation*(2011)

[3] 高野雄一, 横山晶一:語のグループ化を用いた特許文動詞の自動訳し分け, 情報処理学会第 74 回全国大会 4R-3 (2012)

[4] 高野雄一：語のグループ化を用いた特許文動詞の自動訳し分け，山形大学大学院理工学研究科情報科学専攻修士論文(2012)

[5]（財）日本特許情報機構(Japio)：AAMT / Japio 特許翻訳研究特許情報データベース(2008)

[6] 池原他:日本語語彙大系(CD-ROM 版)，岩波書店(1999)

平成 23 年度 AAMT/Japio 特許翻訳研究会

海　外　調　査　報　告

第 13 回機械翻訳国際会議
(Machine Translation Summit XIII)
及び
第 4 回特許翻訳ワークショップ
(The 4th Workshop on Patent Translation)

平成 24 年 3 月

一般財団法人 日本特許情報機構

※　第 13 回機械翻訳国際会議及び第 4 回特許翻訳ワークショップ関連の記事は、AAMT ジャーナル No.50 より転載されたものです。

シンポジウム報告

MT Summit XIII 参加報告

愛媛大学　二宮 崇
山形大学　横山 晶一

第 13 回機械翻訳国際会議（Machine Translation Summit XIII、以下 MT Summit XIII と略称）は、2011 年 9 月 19～23 日、中国福建省厦門の厦門国家会計学院で開催されました。9 月 19 日にチュートリアル、9 月 20 日～22 日に本会議が開催され、9 月 23 日に第 4 回特許翻訳ワークショップが開催されました。会場は、中国に 3 か所（他に北京と上海）ある国立の会計学院で、日本での会計大学院に当たります。設立は 2002 年で、会計士や会社の会計担当重役を養成する機関として、広大な敷地（33.5 万㎡） の中に非常に整った設備（建物面積 7.88 万㎡）が置かれています。参加者は 160 名を超え、地域別ではアジアが最も多く 112 名、アメリカが 23 名、ヨーロッパが 25 名、国別では、中国国内が最も多く 58 名、次いで日本 34 名、アメリカ合衆国 22 名、アイルランドと韓国各 6 名の参加者が集まりました。

私は、20 日の本会議から参加したため、残念ながらチュートリアルには参加できなかったのですが、非常に魅力的なチュートリアルが 4 件ありましたので、それらについて（概要からの抜粋で）簡単に紹介いたします。19 日の午前中は Dekai Wu による「Syntactic SMT and Semantic SMT」の講演が行われ、従来の SMT における統語論(や意味論)の表層的な混合化ではなく、深い理論的なモデルに基づく混合化について解説が行われました。統語論については、様々な transduction grammar の解説を行い、inversion transduction grammar を実現するための足がかりとしての LTG、LITG、PLITG の良さを説明し、意味論については、従来の word sense disambiguation に対する phrase sense disambiguation の優位性について解説がありました。午前中のもう一つのチュートリアル講演は、Mirko Plitt による「Productive Use of MT in Localization」という題目の講演で、ローカル化のための翻訳作業において、翻訳メモリの使用や機械翻訳の後編集による翻訳作業の実態について解説し、Autodesk 社における機械翻訳の取り組みについて解説がありました。午後には、Yanjun Ma、Yifan He、Josef van Genabith による「From the Confidence Estimation of Machine Translation to the Integration of MT and Translation Memory」の講演があり、後編集を想定した、翻訳メモリシステムと機械翻訳システムの統合について解説がありました。同じく午後の Alon Lavie による「Evaluating the Output of Machine Translation Systems」の講演では機械翻訳の人手による評価法と自動評価法を含んだ機械翻訳の評価法一般について解説がありました。

MT Summit XIII の一般講演は 110 件の応募論文のうち 73 件が採択され、うち 55 件が口頭発表、18 件がポスターでの発表でした。内訳は、研究論文が 90 件の応募に対して 62 件採択、ユーザースタディが 14 件の応募中 6 件採択、システムプレゼンテーションが 6 件中 5 件採択になります。一般講演は、3 つのパラレルセッションで行われ、統計的機械翻訳（SMT）に関するモデルや学習の研究、翻訳の精度評価、文法理論、前処理、分野適応、翻訳支援など多岐にわたるテーマの発表がありました。

本会議初日の一般講演では、学習、機械翻訳のための前処理、音声翻訳の 3 セッションによる口頭発表がありました。学習のセッションでは SMT における MERT (Minimum Error Rate Training)パラダイムにおける不安定性を解消する研究が 2 件あり、パラメータ平均化による MERT 最適化の平滑化の研究、BLEU に対する相関係数も目的関数とし

て組み込んだ MERT との混合化の研究が発表されました。同セッションでは、さらに翻訳データのインスタンスに対する重み付けに関する発表がありました。機械翻訳のための前処理のセッションでは、前処理に関する研究や構文解析に関する研究が発表され、動詞と結びつくことによって特別な意味をもつ機能語(particle)まで含めた品詞解析を行うことによって翻訳性能をあげる研究、'do not' における 'do' など対応する語がない時の単語消去に関する研究がありました。

本会議の二日目の午前中のセッションは、学習、機械翻訳を支える技術、翻訳支援の 3 セッションでした。学習のセッションでは追加学習データに対する適応や異なるドメインへの適応など新しいデータに対する適応の研究が発表され、新しく追加される翻訳データに対する効率的なパラメータ更新の研究や、二つのアライメントマトリックスに対するβ分布のパラメータ消去によるベイズ的ドメイン適応の研究、および、統計機械翻訳における能動学習の研究の発表がありました。翻訳支援のセッションでは、参照翻訳を用いなくても計算できる翻訳スコアを用いて、規則に基づく機械翻訳 (RBMT)と SMT を混合したシステムの研究、人間の後編集のプロセスを模した Post-Editing Action (PEA) という枠組みを使って 2 つの異なる MT システムへの後編集を試みた研究、後編集すべき語を翻訳メモリを用いて同定する研究の発表がありました。

二日目の午前は続いて、モデル、言語学的知識に基づく機械翻訳、ユーザースタディの 3 セッションによる口頭発表が行われました。言語学的知識に基づく機械翻訳のセッションでは、中英のアラインメントされたパラレルコーパスを用いて、中国語の単複形を自動的に推定する研究の発表がありましたが、この次の発表がキャンセルになったこともあり、単複をめぐり、会場で様々な討論が行われました。ユーザースタディのセッションでは、Adobe などの企業において機械翻訳がどのように用いられているかについて解説が行われました。

二日目の午後は、モデル、ドメイン適応、マルチパス翻訳の 3 セッションがありました。マルチパス翻訳のセッションでは、語順を入れ替える前処理の規則を依存構造の解析結果から学習する研究、Phrase-Based SMT (PBSMT) において PBSMT を後処理に用いるパイプライン処理の新しい結合方法の研究、英日の翻訳にのみ適用可能な磯崎らの語順入れ替えの前処理を日英の翻訳にも適用可能とする研究が発表されました。その後、ポスター発表やシステムプレゼンテーションが行われました。

三日目の午前中は、モデル、コーパス、機械翻訳のための文法理論の 3 つのパラレルセッションにおける口頭発表と、それに続いてポスター発表とシステムプレゼンテーションがありました。機械翻訳のための文法理論のセッションでは、Dekai Wu による発表があり、彼らが提案した Linear Transduction Grammar が、ある範囲内では、いろいろな言語現象を合理的に説明できるという主張を、このセッションの発表が 2 件だったためか、1 時間にわたりチュートリアル的に発表していたのが印象的でした。その後、同セッションでは、制約付き同期文法の研究の発表が行われました。

三日目の午後には、評価、システム組合せ、ユーザースタディの 3 セッションがありました。システム組合せのセッションでは、うまく翻訳できない原言語の文を言い換えの技術で変換する研究、粒度の細かい RBMT と SMT の混合化の研究、PBSMT における句テーブルを浅い RBMT から得られる辞書で増強する研究、N-best を利用するシステムでの解析誤りの伝播を防ぐためのハイパーグラフに基づく学習の研究発表がありました。

全体的には機械翻訳に対する性能向上の研究が多かったと思いますが、今回の会議では、特許翻訳を中心に、実社会における機械翻訳の利用法や環境整備に関する話を多く聞いたのが非常に印象的でした。現在すでに多くの企業で機械翻訳が実用的に用いられていて、今後、益々必要とされるということではないかと思います。

シンポジウム報告

MT Summit XIII 参加報告

静岡大学
綱川 隆司

2011 年 9 月 19 日（月）～23 日（金）、中国・厦門（アモイ）の厦門国家会計学院（XNAI）において第 13 回機械翻訳サミットが開催されました。本会議は 20 日～22 日の三日間にかけて行われ、19 日にチュートリアル、23 日に併設ワークショップとして第 4 回特許翻訳ワークショップが開かれました。機械翻訳サミットは隔年で開催されており、今回は AAMT の主催、中国・中国語情報学会(CIPS) および厦門大学の後援により中国では初めての開催となりました。

中国の 58 名を筆頭に、日本、アメリカ、欧州等世界各国からの参加者は 160 名を超えました。応募論文 110 編のうち 73 編が採択され、うち 55 件について口頭発表、18 件についてポスター発表が行われました。また、論文の種別の内訳は研究論文 62 件、ユーザスタディー14 件、システムプレゼンテーション 6 件でした。近年広く研究が進められている統計的機械翻訳のモデリングや訓練に関する研究の他、翻訳の精度評価、文法理論、コーパス、翻訳支援、音声翻訳等の多様なテーマにおける発表がありました。

私は 20 日の本会議から参加し、まず董振東教授（中国科学院）から基調講演がありました。翻訳に対する需要が大きいなか、現存する機械翻訳システムと実際の需要とのギャップを埋めるための人間中心の機械翻訳をテーマとして講演をされました。機械翻訳技術のうち人間の翻訳支援に必要な部分を分解し、システム側の知識（辞書、翻訳メモリ等）とユーザ側の知識（個人の語学能力、校正データ等）を統合して一つのプラットフォームとして組み上げることで、翻訳システムと翻訳のための知識を互いに洗練していく流れをつくるべきとの議論をされ、またその一例として格微軟件（Ge-soft）により開発された協調翻訳プラットフォーム、およびそれを用いた比較実験による翻訳能率の向上について紹介されました。

引き続き、Mike Dillinger 氏（AMTA 副会長、TOPs Globalization Consulting 社長）から、“MT Everywhere: Next Steps”と題して招待講演がありました。機械翻訳システムの普及にもかかわらず、それほど多くの企業が機械翻訳を使いたがらない理由は何か。それは機械翻訳システムの価格が高いこと、職業翻訳者による抵抗、あるいは翻訳性能の不完全さのためではなく、機械翻訳そのものに対する理解が不足しているためであり、我々はもっとユーザに機械翻訳を理解してもらうための努力をする必要があると説いています。例えば機械翻訳とは何かと問われた時に、「翻訳ソフトです」と答えると、ユーザは「人間みたいに翻訳してくれるソフトで、翻訳者や通訳の代わりになってくれるだろう」と期待します。そして実際使ってみると、「これを使ったら我が社の翻訳部門をリストラすべきか」といった誤解を生んだり、あるいは、「全然翻訳になってないじゃないか」といったような答えが返ってきたりしてしまいます。それに対して、機械翻訳とは「翻訳を能率化するソフトです」と答えれば、興味を持つユーザが増え、かつ期待した結果が得られるはずだということになります。また、異なる志向のユーザに対して個々に対応すべきであることや、機械翻訳システムを翻訳そのもののために使うのではなくユーザが求めるタスクに対して開発していくことが重要であると述べています。

この後、特許翻訳に関する特別セッションがあり、辻井潤一教授（マイクロソフト・リサーチ・アジア）から序論として特許機械翻訳に関する解説がなされました。特許機械翻訳を行う手段として大規

模並列コーパスの利用や統計的機械翻訳がある中、課題として大量の専門用語の扱い、特許文にありがちな長文の扱いの他、各国の特許を扱う機関の間での協力も必要であるとの議論を展開しました。また、近年の NTCIR において特許翻訳タスクが開始され、特許向けの機械翻訳技術の開発が行われていることの紹介もされました。

この後、チェ・ユチョン氏（韓国特許情報院(KIPI)）と Bruno Pouliquen 氏（世界知的所有権機関（WIPO））からそれぞれ招待講演がありました。チェ氏からは現状での韓国特許庁における機械翻訳の応用例として日韓、韓英および英韓の機械翻訳を用意し、機械翻訳を統合した特許検索システムの紹介と今後の取り組みについて述べられました。また Pouliquen 氏からは利用可能な資源として、英仏の並列特許出願文書コーパスである COPPA、9 か国語で利用可能な言語横断検索システムである CLIR、および WIPO の機械翻訳システムである TAPTA の紹介をされました。

本会議の招待講演は 22 日の最後にもう一件、Hans Uszkoreit 教授（DFKI／ザールラント大学）から、“Strategic MT Research in Europe: Themes, Approaches, Results and Plans”と題して行われました。Uszkoreit 教授は欧州における機械翻訳の指導的研究者であり、EuroMatrix、EuroMatrixPlus、TaraXÜ や META-NET といった様々なプロジェクトに参画しています。これらのプロジェクトの概観の他、現状の研究の進展と課題として、様々な面からの研究が進み新たなアプローチが提案されていく中で、カバレッジの低さや翻訳性能の向上速度の遅さ、機械翻訳の応用範囲の広さに対応するほどの資金が得られていない点等を指摘しています。また、機械翻訳結果のうちよく現れる誤りを含むもの、比較的修正が容易なもの、およびそれ以外の三つの方向に分けて研究を行うという提案や、欧州では外向きおよび欧州内での翻訳需要が比較的大きい点を述べて締めくくりました。

本会議の最後では、東京工科大学の飯田仁教授に IAMT Award of Honor が贈られ、IAMT の次期会長 Andy Way 教授の挨拶と次回機械翻訳サミットの開催地ニースの紹介を行い閉会しました。

23 日には第 4 回特許翻訳ワークショップが開かれ、およそ 70 名の参加者を数え特許翻訳に対する関心の高さを伺わせました。午前中は欧州特許庁(EPO)の Bertrand Le Chapelain 氏、中国特許情報センター（CPIC）の蒋宏飛氏、および日本特許庁の山本英一氏から各国での特許に対する機械翻訳への取り組みに関する招待講演が行われました。欧州特許庁では欧州の各言語から英仏独の各言語への翻訳が要求されており、翻訳システムの要素としてコーパス集、翻訳性能の評価、および翻訳ゲートウェイについて解説されました。中国では特許出願数の増大に伴い機械翻訳の活用を行っており、翻訳支援システムや翻訳システムを組み込んだ特許検索の紹介や、中英・英中に加え日中翻訳も開発中であることを述べました。また日本からは機械翻訳を利用したサービスとして、公開特許公報英文抄録(PAJ)や特許電子図書館（IPDL）の他、高度産業財産ネットワーク（AIPN）における日英機械翻訳の利用、日英・日中機械翻訳研究の促進等を紹介し、将来の展望として中国語や韓国語等での言語横断検索や概念検索のシステム作りを行っていることが紹介されました。

一般講演は 7 件あり、統計的機械翻訳の特許翻訳への適用やその改善、訳し分けへの対処、日本語機能表現の曖昧性の解決、特許翻訳における単語アラインメント、および欧州の機械翻訳プロジェクト MOLTO の特許翻訳への応用についての発表が行われました。またワークショップの最後にはパネル討論が行われ、将来の特許機械翻訳に関する取り組みについて、特にアジア言語の扱いに焦点を置いた議論が交わされました。

全体を通して機械翻訳そのものの改善の話題が中心ではありますが、機械翻訳技術の社会への展開や環境整備などの応用面の重要性が強調されていたことが私としては印象に残った会議でした。

シンポジウム報告

MT Summit XIII における「特許翻訳 WS」等に関する報告

(財)日本特許情報機構 特許情報研究所
調査研究部長 森藤 淳志

１．はじめに

筆者は、平成２２年１０月から縁あって（財）日本特許情報機構（Japio）に在籍し、AAMT/Japio 特許翻訳研究会を通じて、AAMT の活動に関与することができる機会に恵まれています。この度、MT サミット XIII(Xiamen)における特許翻訳に関するワークショップ(WS)と特別セッションのプログラム委員(Program committee)として関わり、MT サミットに初めて参加しました。

AAMT において筆者は新参者の部類に入ると考えられますので、本稿では、まず筆者と MT サミットとの関わりについて簡単に触れたいと思います。また、第４回特許翻訳 WS 等の詳細な報告は他の参加者からなされるようですので、筆者からは今回の MT サミットについて補完的な報告をいたします。

２．筆者と MT サミットの関わりなど

筆者と MT サミットの関わりは 2007 年に開催された MT サミット XI（Copenhagen）に遡ります。

筆者は当時、日本特許庁にて、特許公報の日→英機械翻訳(MT)を一般に提供する「特許電子図書館(IPDL)・英語版」や拒絶理由通知などの審査書類の日→英 MT を他国審査官に提供する「高度産業財産権ネットワーク(AIPN)」（2011 年８月現在４８の国・地域へ提供）を担当していました。

この頃に、第２回特許翻訳ワークショップでの講演の機会を頂戴し、派遣者の決定や講演内容の推敲を通じて、陰ながら MT サミットに関わりました（私の配下の遠山敬彦 が参加・講演しました)。

３．日本特許庁におけるＭＴ活用と今回の講演

日本特許庁は１９９９年に IPDL を立ち上げ、その後まもなく、日→英 MT を活用したサービスの提供を開始し、さらに、2004 年 10 月から AIPN の運用を開始しています。特許情報に機械翻訳を適用して、ネットでサービス提供することは、当時、他庁のサービスと比較して先駆的な試みでした。

しかしながら、今回の MT サミットで講演された、各庁（日本特許庁、世界知的所有権機関、欧州特許庁、韓国特許庁、中国特許庁）における機械翻訳の活用状況に関する内容を見る限り、その後の世界知的所有権機関など他庁における MT 活用の取り組みの状況には目を見張るものがあります。例えば、韓国特許庁から、審査官と一般公衆に対して、日本特許公報の日→韓 MT サービスが提供され、欧州特許庁からは多言語間の MT サービスが提供されています。日本特許庁の上記 MT サービスは、日→英 MT の精度改善にとどまっていて、日本企業等ユーザ向けに外国語→日本語 MT サービスは未だ実現されていません（民間プロバイダが一部実施)。

４．日本特許庁の国際知財戦略と今回の講演

経済産業大臣の諮問に応じて、経済及び産業の発展に関する重要事項を調査審議する産業構造審議会の下部には、特許等の産業財産権に関する政策審議を行う場として、知的財産政策部会が設置されています。さる平成２３年７月１９日に第 16 回産業構造審議会 知的財産政策部会が開催され、「国際知財戦略（Global IP Initiative）～国際的な知的財産のインフラ整備に向けた具体策～」（以下、「国際知財戦略」）が審議されました。

この国際知財戦略の資料中で、「日本語・英語以外の特許文献、特に急増する中国文献に対し、企業・特許庁ともに戦略の転換が必要」であること、「中国において、無審査登録の実用新案権に基づき、賠償を求められる事例が出現。分類や翻訳の整

備により、中国文献などを容易に把握できるように」することの必要性が報告されました。同資料中で、こうした現状を踏まえ、中→日、韓→日 MT 機能を備えた外国特許文献検索システムの整備を行うという方向性が示されています。

この点、今回の MT サミットにおける日本特許庁山本氏のプレゼンテーションでは「The JPO’s New Search System」というスライドにおいて、「Cross-lingual search system (Chinese, Korean, etc.)」という事項が含まれていることから、日本特許庁は上記審議会での審議結果を実施に移す意向があることが伺うことができます。

Japio での研究活動を通じて、筆者も、日本のグローバル企業の知的財産部の方と意見交換をすることがあります。日本を代表するグローバル企業が、中国文献の未曾有の増加や訴訟件数の急増（訴訟大国・米国の 2 倍に到達）に対して苦慮している現状を垣間見るにつけ、日本特許庁の上記施策が実現されること、その際には、日本における中日 MT の研究成果が活用され、高品質な MT 結果が得られるようになることを希求します。

５．本 MT サミットにおける Japio の貢献

Japio は、毎年 11 月に主催する特許・情報フェア＆コンファレンス（東京）等において、世界知的所有権機関や欧州特許庁などから特許情報専門家を招聘したり、別途、研修生を受け入れたりするなど、海外の特許庁とのネットワークを有しています。このネットワークを活用し、また、日本特許庁の力添えも頂戴しながら、Japio は、特許翻訳特別セッションと第４回特許翻訳ＷＳへの各国特許庁から招待講演者の招へいを担当しました。

また、ＷＳでは、Opening Addressを当機構の専務理事守屋が担当し、その中で、Japioの機械翻訳への取り組み（Japioコーパス、AAMT/Japio研究会）に加え、多言語化に向けて中国語対応に着手したことを報告しています。

６．開催地アモイ；観光地？ストイック？

厦門（アモイ：Xiamen）は、中国福建省の南東部、台湾の対岸に位置する島で、南西に隣接するコロンス島は、歴史的建築物が数多くあり、観光スポットとして多くの人が訪れます。他方、今回の会場となった Xiamen National Accounting Institute（XNAI）は、主に中国国内の会計担当者のための研修施設であり、観光スポットから隔離されたストイックな環境であることから、筆者は MT サミットに集中できました。下図：XNAI の全景図

７．むすび： 謝辞と今後

今回の特許翻訳ＷＳ等は、成功裏に終了し、各国特許庁からの出席者にとっても有益であったと思います。この特許翻訳ＷＳ等の企画段階から当日の運営に至るまで、辻井委員長以下 AAMT/Japio 特許翻訳研究会の全員で対応してきました。特に、chair を勤めていただいた横山先生と co-chair の江原先生による、各国特許庁からの招待講演者に対する紳士的な対応も成功の一因と思います。委員長と両氏を始め関係者の皆様に深く御礼申し上げます。

最後に一言付記します。昨今の記録的な円高などもあり、日本の企業の海外展開はさらに加速することが予想されます。直近では、中国特許文献が注目されていますが、今後は、ロシア語などの特許文献への対応が必要となると予想されます。まずは、中国語での機械翻訳を成功させ、次のターゲット言語にも速やかに展開していくことが必要となると思います。今後の非英語圏 MT の発展に期待します。

シンポジウム報告

MT Summit XIII Technology Showcase 報告

東芝ソリューション株式会社
熊野 明

MT Summit XIII の本会議が開催された 9 月 20 日から 22 日までの 3 日間、機械翻訳に関連する技術展示を行う Technology Showcase が開かれた。論文発表会場 XNAI 会議棟の 1 部屋で、7 団体がそれぞれの技術や製品を紹介した。筆者は、自社の技術・製品の紹介のために 3 日間参加した。

学会参加者には、技術文書のローカライゼーション、海外の特許文書の検索など、機械翻訳の実運用に関心のある人が多く、論文発表や講演セッションの休憩中に、頻繁に展示会場を訪れていた。参加者の関心は、製品の仕様だけでなく、しばしば翻訳エンジンに関する技術的な項目にも及び、展示説明者と積極的な議論を行なっていた。

以下に各団体の展示概要を紹介する。(ABC 順)

○ Baidu（百度）

中国最大の検索サイトを運営する企業。Web から収集した対訳テキストを利用し、SMT(Statistical MT)をメインエンジンとした機械翻訳を開発し、ポータルサイトで公開している。

○ ETRI（Electronics and Telecommunication Research Institute）

韓国の研究機関であり、韓国語と英語、韓国語と中国語の双方向の機械翻訳技術を開発している。2005 年には特許文書を対象とした韓国語から英語への機械翻訳を、翌 2006 年には英語から韓国語への機械翻訳を製品化し、KIPO（韓国特許庁）にも技術を提供している。現在は技術文書だけでなく、話し言葉に対する機械翻訳の開発にも取組んでいる。

○ Fujitsu Research and Development Center（富士通研究開発中心）

富士通が中国に設立した研究開発会社。機械翻訳の技術を利用して、中国語の特許を日本語で検索するなど、言語横断検索の技術を紹介した。

○ Microsoft Research

検索サービス Bing で公開している機械翻訳は SMT のエンジンを利用しており、Microsoft Office のアプリからも利用できる。CTF（Collaborative Translation Framework）は、3 種類のメンバーの分業によって文書翻訳を行なう体系。複数の User が翻訳した結果を、Moderator が確認、修正して承認する。Owner は全体を管理し、翻訳文書を完成する。

○ NICT

Web を利用することにより、複数の人の手で文書全体の翻訳を行なう“みんなの翻訳”のプロジェクトを紹介した。

○ Shenyang Global Envoy Software（瀋陽格微軟件公司）

自社の機械翻訳を活用した翻訳ビジネスを行っている中国企業。航空機マニュアルのような大規模なドキュメントに対し、専門のオペレータを集めたプロジェクトを組織化して、翻訳を行なう。前処理では、原文書から専門用語を抽出して訳語を決定する。後処理では、類似の原文に対する訳文の表現を統一し、全体で均質な訳文を実現している。

○ 東芝ソリューション

日本の企業として唯一の出展。日本語と英語、日本語と中国語の双方向の機械翻訳を、企業向けサーバシステム“The 翻訳エンタープライズ”、クラウドサービス Eiplaza/MT で提供している。パッケージソフト“The 翻訳プロフェッショナル”（日本語・英語双方向）では、SDL Trados Studio との連携機能により、プロの翻訳家に向けた翻訳環境を実現した。翻訳エンジンは、RBMT(Rule-based MT)とEBMT(Example-based MT)を組合せたものである。

複数の企業の社内文書ローカライゼーション担当者からは、SDL Trados Studio との連携に関して具体的な質問を受け、関心の高さが明らかになった。

今回は残念ながら 7 団体の参加にとどまったが、2 年後にはより多くの団体の出展を期待したい。

平成23年度 AAMT/Japio特許翻訳研究会

海　外　研　修　報　告

南カリフォルニア大学情報科学研究所
（USC/ISI）研修報告

平成24年3月

一般財団法人 日本特許情報機構

研修報告～南カリフォルニア大学情報科学研究所(USC/ISI)

北海学園大学　越前谷 博

1　南カリフォルニア大学情報科学研究所（USC/ISI）について

2011 年 4 月から 2012 年 3 月までの 1 年間、本学の在外研修制度を利用し、アメリカ、カリフォルニア州のロサンゼルス郊外にある南カリフォルニア大学情報科学研究所（University of Southern California/Information Sciences Institute：以下、USC/ISI と呼ぶ)にて研究を行う機会に恵まれたため、そこでの研修生活について報告させていただく。

USC/ISI は USC のメインキャンパスから約 20km 程度西側に位置するマリナ・デル・レイに、1972 年に創設された USC の研究所の一つである。筆者はこの USC/ISI の Intelligent Systems Division の Visiting Scholar として籍を置かせていただいた。Intelligent Systems Division は Natural Language Technologies、Information Integration、そして、Knowledge Technologies の 3 つのグループから構成されており、人工知能分野全般を網羅している。また、これらのグループは独立しているわけではなく、相互に交流可能な環境が整っている。筆者は Natural Language Technologies の Deputy Director である Eduard Hovy 氏を通じて USC/ISI での研究生活を送ることができた。Natural Language Technologies には Eduard Hovy 氏を始め、Kevin Knight 氏、Daniel Marcu 氏、Jerry Hobbs 氏、Gully Burns 氏、Hans Chalupsky 氏といった第一線で活躍する著明な研究者達が名を連ね、この方々が Project Leader を務めている。また、世界各国の学生を含む多くの優秀な研究者がここ USC/ISI に籍を置き、研究活動に励んでいる。その結果、USC/ISI から優秀な研究成果が発信され、USC/ISI は現在の自然言語処理分野はもとより、人工知能分野をリードし続けている。

2　USC/ISI での研究生活

USC/ISI の研究者はマリナ・デル・レイにある Marina Towers という 12 階建てのツインタワーの一つであるサウスタワーの 4 階と 9 階の部屋を研究室として利用している。筆者には Eduard Hovy 氏のいる同じ 4 階の個室が研究室として与えられた。他の Visiting Scholar にも個室が与えられており、これはかなり恵まれた環境といえる。

USC/ISI は多くの研究者との交流の場として最高の環境を有している。具体的には、不定期ではあるがほぼ毎週、AI セミナーと NL セミナーの両セミナーが開催され、発表者は USC/ISI の研究者をはじめ、アメリカ、ヨーロッパから訪れる外部の大学や研究機関の研究者である。そして、その内容は多岐に渡っている。NL セミナーでの発表内容は自然言語処理分野に限定されているが、AI セミナーは自然言語処理だけでなく、バイオインフォマティックスや地理情報科学など様々な研究テーマに関する内容となっている。その中から筆者が聴講したセミナーをいくつか紹介させていただく。

また、この研修生活で著者が取り組んだ研究内容についても簡単に述べさせていただく。

2.1 セミナー紹介

(1) The Copiale Cipher ：AI セミナー（2011 年 7 月 8 日）

Kevin Knight 氏(USC/ISI)：Kevin Knight 氏は USC/ISI の Project Leader のお一人であり、統計翻訳研究の第一人者である。250 年前に東ベルリンで発見され、100 ページ以上、数千個以上の文字で書かれた暗号文書として知られている「Copiale 暗号」を、統計翻訳でも利用されている EM アルゴリズムに基づく手法によりその解読に成功した。そして、「Copiale 暗号」には 18 世紀のドイツの秘密結社の儀式と政治的な見解に関する記述があることを発見した。また、Kevin Knight 氏はヴォイニッチ手稿やゾディアックの暗号文の解読にも既に取り組んでいる。

(2) Overcoming Information Overload in Navy Chat：NL セミナー（2011 年 8 月 5 日）

Dave Uthus 氏(Naval Research Laboratory)：インターネット・リレー・チャットのような伝達速度を重視したチャットは軍の分野でも重要な役割を果たしている。アメリカ海軍においては、モニターチェックを行う「watchstanders」と呼ばれる人たちが複数のチャットから同時に戦術に関する会話のモニタリングを行っているが、その際の膨大な情報量が問題となっている。この問題に対して、緊急性の高い重要なメッセージのみを探知し、かつ、要約による情報圧縮を行うことで、その解決を図っている。膨大な軍事力を持つアメリカならではのタスクであり、大変興味深かった。

(3) HyTER : Meaning-Equivalent Semantics for Understanding, Generation, Translation, and Evaluation：AI セミナー（2012 年 2 月 3 日）

Daniel Marcu 氏(USC/ISI)：USC/ISI の Project Leader のお一人である Daniel Marcu 氏による機械翻訳の自動評価に関する講演である。語彙単位でアノテートされた翻訳文と参照訳を HyTER networks と呼ぶネットワークで表現し、これらの間でレーベンシュタイン距離の最小値を求めることで翻訳文を評価する。アノテーションは人手で行われる。性能評価実験では、BLEU や METEOR よりも高い精度を持つ HTER との相関が得られた。また、HyTER networks は人間の翻訳者の能力を評価するために利用可能である。

(4) Large Scale Syntactic Language Modeling with Treelets：NL セミナー(2012 年 2 月 17 日)

Adam Pauls 氏(University of California, Berkeley)：構文解析のための言語モデルの提案とその評価についての講演である。提案手法では、標準的な n-gram 言語モデルを用いて、パージングされたテキストから頻度を収集することにより自動的にモデルのパラメータを推定する。その際、シングルマシーンを用いて数時間で 100 万以上のトークンからモデルを学習することができる。実験の結果、正データのみの学習にもかかわらず、他手法と比べ、高い性能を示した。

今回は割愛させていただくが、上述したセミナー以外にも様々な興味深い講演を聴講することができた。また、USC/ISI にはバイオインフォマティクスの優秀な研究者が多く在籍しているため、AI セミナーではバイオインフォマティックスに関するセミナーが数多く開催された。このような USC/ISI でのセミナーへの参加は、USC/ISI に籍を置かせていただいたことによる恩恵の一つであり、著者にとっては様々な研究分野の先端技術に触れることができる場として、大変貴重な機会となった。

2.2 USC/ISI での研究内容

USC/ISI での研究生活では、Machine Reading の研究に取り組んだ。しかし、この Machine Reading の解釈は幅広く、アメリカ国防総省国防高等研究計画局（DARPA）では、「エキスパートや知識エンジニアに変わり、自然テキストから知識を直接抽出すること」と位置付けられており、また、Machine Reading 分野の第一人者であるワシントン大学の Oren Etzioni 氏は「人手でタグ付けされたトレーニングデータを用いる従来の教師付き学習とは異なる、教師なし学習による自然テキストの理解である」と位置付けており、幅広い解釈が可能である。

そこで、Eduard Hovy 氏に「我々はどのようにして Machine Reading の研究であるか否かを判断すべきなのか」と直接尋ねてみた。Eduard Hovy 氏はあくまでも個人的な見解であり、他の研究者とは違うかもしれないという前置きをされたうえで「確かに Machine Reading の定義は明確ではないが、私は大規模テキストを対象とした浅い自然テキストの理解ではなく、1 文単位もしくは非常に小さな単位のテキストを対象とした深い理解であると考えている」との回答を下さった。

著者が研修中に取り組んだ研究内容は、インターネット上から収集した自然文を対象に、2 つの arguments の間の relation を自動抽出するものである。最終目的は、大規模な言語データベースの構築ではあるが、その第一歩として、Initial seed 及び Predefined relation の使用を前提とせずに、様々な文から relation を抽出するための研究を行った。また、このようにして抽出される言語知識は機械翻訳の自動評価の研究においても非常に有益であると考えている。これまでに機械翻訳の自動評価手法として IMPACT を提案してきたが、string レベルの情報のみでは翻訳文の自動評価には限界があり、更に評価精度を向上させるためには、高度な言語情報の利用が不可欠である。したがって、研修中に行った研究は、今後も継続して取り組むべき、非常に重要なものと考えている。

3 あとがき

このような大変有意義な研修生活を送れたことは、研究者として非常に幸せなことである。また、一年間ではあるが海外生活を送れたことは大変貴重な経験であった。このような機会を与えていただいた、本学に対して感謝の意を持っていることは言うまでもないが、同時に著者の訪問を快く受諾して下さり、かつ、アメリカで快適な生活を送るために常に気を配って下さった Eduard Hovy 氏に心から感謝の意を表したい。

また、今回の研修が実現した大きな要因として、著者が 2010 年 7 月にスウェーデンで開催された国際会議 ACL で口頭発表を行った際に、セッションチェアーを Eduard Hovy 氏が務められ、そこでお話しする機会を持てたことが挙げられる。その際の発表内容は Patent data を用いた機械翻訳の自動評価手法についてであり、AAMT/Japio 特許翻訳研究会のメンバーとして活動させていただかなければ ACL での発表は実現せず、その結果、今回の USC/ISI での研修生活もまた実現しなかったと考えている。本研究会辻井潤一委員長を始め、本研究会のメンバーの方々にもこの場を借りて感謝の意を表したい。今後は、これまで以上に本研究会に貢献できるよう、研究に精進する所存である。

Memo

平成23年度AAMT/Japio特許翻訳研究会報告書
（機械翻訳及び辞書構築に関する研究及び海外調査）

発 行 日　平成24年3月

発　　行　一般財団法人 日本特許情報機構（Japio）
〒135-0016 東京都江東区東陽4丁目1番7号
佐藤ダイヤビルディング
TEL:(03) 3615-5511　FAX:(03) 3615-5521

編　　集　アジア太平洋機械翻訳協会（AAMT）

印　　刷　株式会社 ナビックス