

4-4 構文情報を直接利用した機械翻訳システム コンビネーション

4-4 *Direct Use of Syntactic Information for Machine Translation System Combination*

渡辺太郎

WATANABE Taro

要旨

現在主流の機械翻訳システムコンビネーションでは、単語単位の近さを基準に、複数の出力のアライメントを計算し、confusion networkなどのグラフ構造で表現する。本稿では、複数の木構造を圧縮し、コンパクトな構文森として表現した、confusion forestを用いた、新しいシステムコンビネーションの手法を提案する。構文森はシステムの出力の構文解析結果か統語論的な一致を計算することで先制される：まず、システムの出力を構文解析する。構文解析木を構成するルールを抽出し、文脈自由文法を獲得する。獲得された文法に基づき、開始記号から始め、非終端記号を書き換えることにより構文森を生成する。新しい翻訳は構文解析森から最適な導出木を求めることにより得られる。統計的機械翻訳ワークショップ (WMT10) における機械翻訳システムコンビネーションタスクにおける実験結果から、従来法に比べほぼ同様な翻訳を得ることができ、仮説を大幅にコンパクトな空間に表現できることを確認した。

The state-of-the-art system combination method for machine translation (MT) is based on confusion networks constructed by aligning hypotheses with regard to word similarities. We introduce a novel system combination framework in which hypotheses are encoded as a confusion forest, a packed forest representing alternative trees. The forest is generated using syntactic consensus among parsed hypotheses: First, MT outputs are parsed. Second, a context free grammar is learned by extracting a set of rules that constitute the parse trees. Third, a packed forest is generated starting from the root symbol of the extracted grammar through non-terminal rewriting. The new hypothesis is produced by searching the best derivation in the forest. Experimental results on the WMT10 system combination shared task yield comparable performance to the conventional confusion network based method with smaller space.

[キーワード]

統計的機械翻訳, システムコンビネーション, 句構造解析木, 森

Statistical machine translation, System output combination, Phrase structure parse tree, Forest

1 はじめに

システムコンビネーションは、音声認識 [1][2] や構文解析 [3] など幅広い分野に応用され、複数のシステムの出力の合意を取り、出力を決定する。現在主流の機械翻訳システムコンビネーション手法は、複数の仮説を表現したコンパクトなグラフ構造である、confusion network に基づ

く [4]。

confusion network は文字列の近さにより構築される。まず、スケルトン、あるいはバックボーンと呼ばれる仮説が1つ選択される。他の仮説は、スケルトンに対してアライメントされ、各辺が別の単語の候補を表したラティス構造を形成する。アライメントの手法として、モデルに基づく手法 [5][6] では統計的な単語アライメント手法を

用いるのに対し、編集距離に基づく手法 [7][8] では、例えば翻訳エラーレート (translation error rate、TER) [9] などの翻訳の評価指標によりアライメントを決定する。新しい翻訳候補は構築されたネットワークから最適な経路を選択することにより生成される。

本稿では、システムの出力の統語論的な近さを利用した、全く新しいシステムコンビネーション手法を提案する。文字列に基づく confusion network を構成するのに対し、複数の木構造を多項式な空間で表現した、圧縮森 [10][11] を生成する。圧縮森、あるいは *confusion forest* は機械翻訳の出力の統語論的な同意を取ることで複数の出力を結合する。confusion forest は以下のように、文法に基づく手法で形成される：まず、機械翻訳システムの出力を句構造解析する。次に、句構造解析木から、その木構造を構成する規則の集合を抽出する。規則の集合に基づき、開始記号から始め、アリー法 [12] により圧縮森を生成する。新しい翻訳候補は、この森から最適な導出木を求めることで選択される。ここで抽出される規則の集合、あるいは文法は、構文解析された木で用いられたものに制限される。最後の森を生成するステップにおける冗長な曖昧性は、例えば、アリー法で用いられる状態表現を用い、各ノードの親、あるいは兄弟ノードを記憶することで、元々の木構造の各ノードに木構造の局所的なコンテキストの情報を残し、曖昧性を解消する。

統計的機械翻訳ワークショップ (WMT10) における、チェコ語およびドイツ語、スペイン語、フランス語の各言語から英語への機械翻訳システ

ムコンビネーションタスク [13] において実験を行った。この結果、2つの言語対に対し、従来法に比べほぼ同様な翻訳を得ることができ、他の言語対についても統計的に有意な改善が見られた。

2 Confusion Network

confusion network に基づく機械翻訳システムコンビネーションは、まず、各出力間のアライメントを計算する。Matusov [5] らは GIZA++ [14] などの統計的単語アライメントモデルによるアライメントを利用して仮説のアライメントを計算している。Sim [8] らは TER [9] など編集距離を利用している。次に、複数の出力の中から、例えば Minimum Bayes Risk などの基準を用いて、単語の並び替えの基準となるスケルトンを選択する。他の仮説は、スケルトンに対するアライメントに基づいて組み合わせられ、ネットワークが形成される。図 1 (a) はスケルトンに対し、複数のシステムの出力をアライメントした例を示している。図 1 (a) から、各エッジに ϵ を含む、単語のラベルが付けられた図 1 (b) のような confusion network が生成される。このようなペア単位のアライメントでは、例えば、「green」が「forest」と対になるエラーのため、単語の繰り返しが多く見られる。逐次的な手法 [15] では、仮説単位のアライメントを計算するのではなく、逐次的に仮説が組み合わせられる confusion network と、各出力とのアライメントを計算している。これにより、例えば図 1 (c) のように、「green trees」と「blue forest」が対となり、ネットワークが

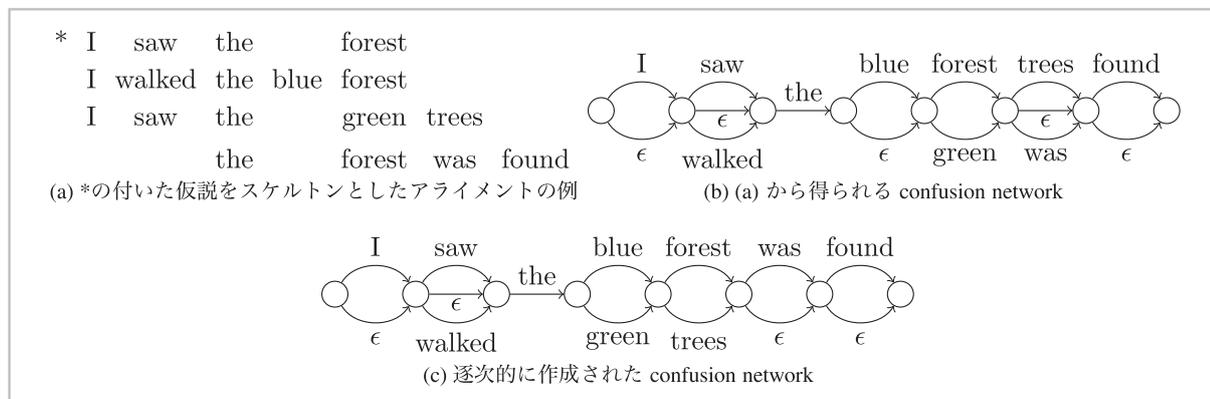


図 1 confusion network の例

形成される。

confusion network に基づく手法は、単語の並びを決定するスケルトンの選択に依存している。例えば、図 1 (a) の例では、最初の 3 つの仮説が能動態であるのに対し、最後の仮説は受動態であり、文法的に大幅に異なる構造を持つ。この結果、図 1 (c) のように極端に長い仮説を生成することが可能となる。これを部分的に解決する手法として、[16] は各出力それぞれがスケルトンとして複数のネットワークを作成し、そのネットワークを結合して大きなネットワークを作成する手法が提案されているが、根本的な解決とはなっていない。

3 Confusion Forest

複数の仮説を confusion network として表現するのに対し、本稿では、confusion forest という複数の構文木をコンパクトに表現した構文森として表現する。統語的な合意は構文解析木の木構造の断片を共有することにより実現する。構文森は構文解析や機械翻訳で使用されるハイパーグラフ (hypergraph) として表現される [17]。具体的には、ハイパーグラフは $\langle V, E \rangle$ という二つ組で表現され、 V, E はそれぞれノードの集合、ハイパーエッジ (hyperedge) の集合とする。 V の各ノードは $X^{@p}$ として表現され、 $X \in N$ は非終端記号であり、 p は各ノードの ID を親からの相対的な位置として表されるアドレスとする [18]。例えば、ルートノードには、 ϵ のアドレスが割り当てられ、 p の最初のノードには、 $p.1$ のアドレスが割り当てられる。各ハイパーエッジ $e \in E$ は文脈自由文法の規則のインスタンスであり、 $\langle head(e), tails(e) \rangle$ という二つ組で表される。推論規則とみなした場合、 $head(e) \in V$ は後件を表したノードであり、 $tails(e) \in V^*$ は前件の集合である。図 1 (a) の仮説に対する構文森の例を図 2 に示す。例えば、 $VP^{@2}$ の 2 つのハイパーエッジ $\langle VP^{@2}, (VBD^{@3}, VP^{@4}) \rangle$ と $\langle VP^{@2}, (VBD^{@2.1}, NP^{@2.2}) \rangle$ により、前者は受動態、後者は能動態という文法的に異なる導出が可能である。

機械翻訳システムの出力が与えられたとき、以下のような文法的な手法により confusion forest

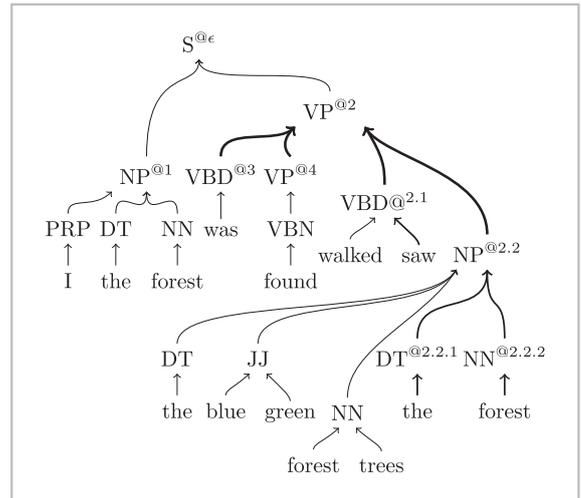


図 2 図 1 (a) を基に作成された confusion forest の例

を生成する。まず、機械翻訳システムの出力を構文解析する。次に、各構文解析木の各ハイパーエッジを文脈自由文法の規則のインスタンスとみなして文法を獲得する。入力文に特化した局所的な文法を基にして、開始記号から始め、各非終端記号を書き換えていくことにより構文森を生成する。新しい仮説は、生成された構文森から最適な導出木を計算することにより得られる。

3.1 ルールの獲得

ルール獲得時に、構文木の各ノードに割り当てられている非終端記号に元々の木構造の形を符号化することにより、ルールの曖昧性を減らす。まず、水平的な Markovization [19] により、各ノードの非終端記号にその左右の兄弟ノードの非終端記号を符号化する。例えば、図 3 (a) は「I saw the forest」に対する解析木の例を示す。図 3 (b) では、図 3 (a) のノード $VP^{@2}$ をルートとした木構造に対するラベルの書き換え例を示す。例えば $NP^{@2.2}$ に対して、その左にある兄弟ノード $VBD^{@2.1}$ のラベルを組み合わせ、 \bullet で元々のノードのラベルの位置を示す。続いて、垂直的な Markovization [19] により、親ノードのラベルを組み合わせる。図 3 (c) では、 $@.2.2$ のノードがその親である $@.2$ のノードのラベルと組み合わせられ、 $(NP: \bullet VP + VBD: \bullet NP)$ のようなラベルが得られる。このようにラベルの書き換えを行った後に、各ハイパーエッジをルールとみなし

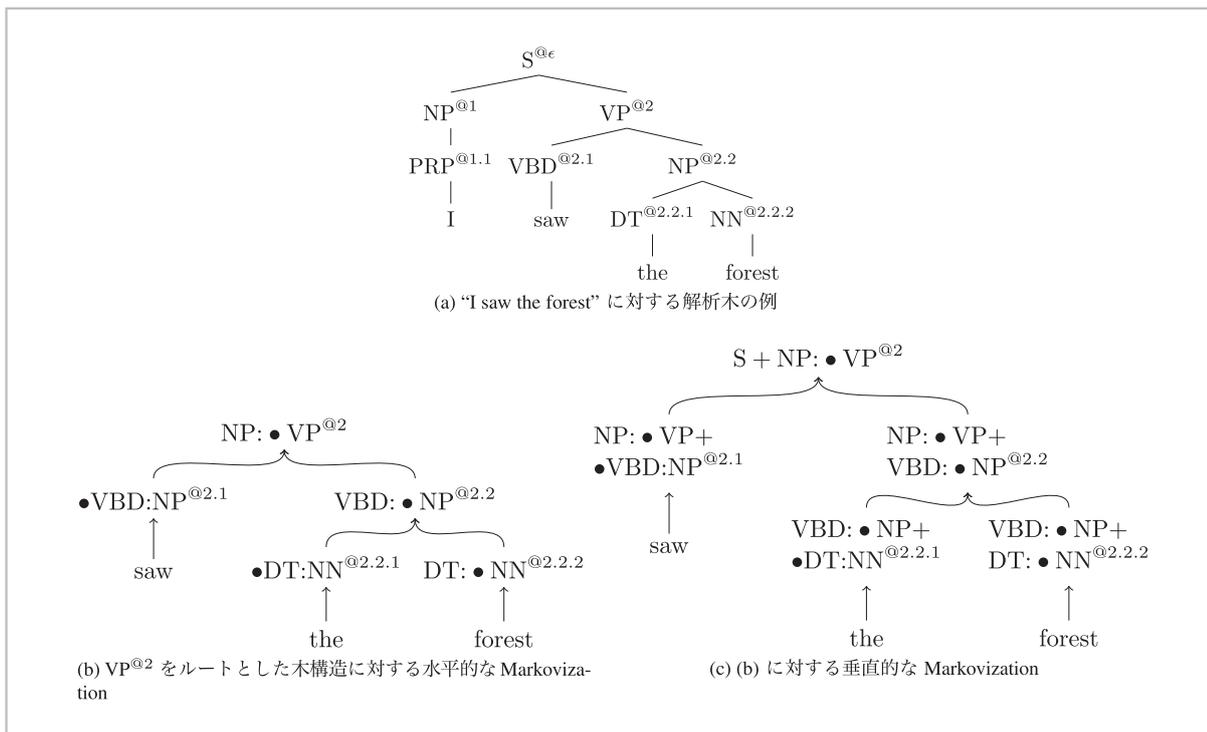


図3 水平、垂直 Markovization によるラベルの書き換え例

て文法を学習する。

各ノードで表された木のコンテキストの情報は水平および垂直的 Markovization により制限される [19]。垂直的オーダ v により最大 v の親ノードの数を記憶する。同様に、水平的オーダ h により、 \bullet の左右それぞれ最大 h 個の兄弟ノードを記憶する。

水平的および垂直的な Markovization のオーダに制限を設けない場合、元々の構文解析木を記憶することになり、ここから得られる文法で生成される confusion forest は、構文解析結果の和集合を取ることになる。水平的なオーダを少なくすることにより、兄弟ノードの情報を捨てることとなり、より木構造の並び替えを許すことになる。同様に、垂直的なオーダを制約することで、各ノードに達するまでのパスを無視することで、深い森を生成することになる。

3.2 構文森の生成

3.1 で得られた文法に対して、アーリー法 [12] を適用し構文森を生成する。推論規則 [20] として表現された生成アルゴリズムを図4に示す。X ∈ N は非終端記号とし、x ∈ T を終端記号と

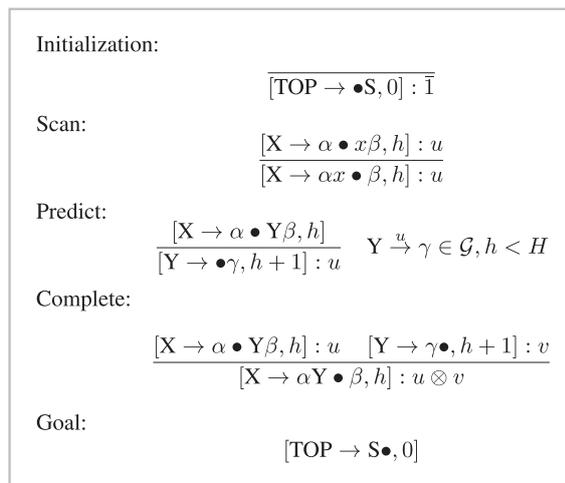


図4 アーリー法生成アルゴリズムの推論規則

する。a と β、γ は終端記号、非終端記号の記号列 (T ∪ N)* であり、u と v は、各項目に割り当てられる重みである。

一般的なアーリー法とは異なり、各非終端記号に割り当てられるスパンの情報は無視され、各導出に対する高さを h として保持する。Scan ステップは必ず成功し、このため、深い構文森が生成される。この深さは、Predict ステップにおける $h < H$ により制限される。本稿では、 H は構

文解析されたシステムの出力のうち最大の深さの1.5倍としている。

3.3 構文森のリスコア

構文森 F から、構文森に基づく k -best 構文解析アルゴリズム [21] を使用して、 k -best の導出 d を求める。全ての可能な導出 D から最適な導出 \hat{d} を求めるため、複数の素性を線形結合した目的関数を使用する。

$$\hat{d} = \arg \max_{d \in D} \mathbf{w}^\top \cdot \mathbf{h}(d, F) \quad (1)$$

ここで、 $\mathbf{h}(d, F)$ は素性の集合であり、 \mathbf{w} により重み付けされる。Cube Pruning により、N-gram 言語モデルなどの非局所的な素性との近似的な結合を行う [22][23]。そして、 k -best 導出には、Huang と Chiang のアルゴリズム 3 を用いる [21]。

4 実験

4.1 設定

WMT10 における、チェコ語およびドイツ語、スペイン語、フランス語の各言語から英語への機械翻訳システムコンビネーションタスクにて実験を行った [13]。データを表 1 に示す。各システムの出力は Stanford parser [19] にて構文解析し、全ての単語を小文字へと変換した。

confusion forest に基づくシステムコンビネーションを、ハイパーグラフツールキット *cicada* で実装した。*cicada* ツールキットは、汎用的な重み付け演繹論理システム [24] に基づき、同期文脈自由文法に基づく機械翻訳システム [22] を実現している。**3** で述べたように、各翻訳システムの出力の句構造解析の結果をハイパーグラフの集合を入力として、ルールを獲得し、新たに、構文森を生成する。ベースラインとして、*cicada* で実装された confusion network に基づくシステムコンビネーション手法 (**2** 参照) は、TER に基づき、各仮説のアライメントを逐次的に行い、かつ、複数のスケルトンに基づいて作られたネットワークを1つのネットワークとして表現している [15]。エプシロン遷移を取り除いた後、ネットワークを $S \rightarrow X$, $S \rightarrow SX$, $X \rightarrow x$ から成る文法で構文解析し、森へと変換した。**3.3** での森に

表 1 WMT10 システムコンビネーションのデータ

	cz-en	de-en	es-en	fr-en	
システム数	6	16	8	14	
平均文字数	tune	10.6K	10.9K	10.9K	11.0K
	test	50.5K	52.1K	52.1K	52.4K
文数	tune	455			
	test	2,034			

対するアルゴリズムにより、変換された森から新たに k -best の翻訳候補を導出している。

4.2 素性

式 (1) における素性の重みベクトル \mathbf{w} はハイパーグラフエラー最小化学習 (MERT) により学習する [25]。

以下のように、複数の素性を用いた: 3つのコーパスから5-gram 言語モデル $h_{lm}^i(d)$ を学習: English Gigaword 第4版^{*1}、French-English 10⁹ コーパスおよび news commentary English data の英語側^{*2}。 $h_l(d)$ と $h_e(d)$ は、それぞれ、 d における非終端記号及びハイパーエッジの数を数える。 M 翻訳システムの各システム m に対する信頼度 $h_s^m(d)$ として、 d で使われているルールのうち、 m のシステムの仮説から得られたルールの数を用いる [26]。

M 個のシステムの出力 e_1, \dots, e_M が与えられた時、[27] に基づき、BLEU [28] に基づく相関による素性 $h_b^m(d)$ を用い、 e_m を参照訳として、 d の BLEU を計算する。

$$h_b^m(d) = BP(e, e_m) \cdot \exp\left(\frac{1}{4} \sum_{n=1}^4 \log \rho_n(e, e_m)\right)$$

ここで、 $e = \text{yield}(d)$ は、 d から求められる非終端記号列とし、 $BP(\cdot)$ および $\rho_n(\cdot)$ は、それぞれ簡潔ペナルティ (brevity penalty) および N-gram 適合率とする。ここで、コンパクトな状態表現 [29] により、制約のない (unclipped) N-gram の数 [30] を使用した。

*1 LDC catalog No. LDC2009T13

*2 これらのデータは、<http://www.statmt.org/wmt10/> から得られる。

confusion network によるベースラインシステムは、さらに、 m 番目のシステムの出力をスケルトンとして生成されたネットワークの編集距離数をそのネットワークのノード数でスケーリングしたペナルティ素性 $h_p(m)$ を用いる [16]。

4.3 結果

異なる水平的 ($h = 1, 2, \infty$)、垂直的 ($v = 4, 5, \infty$) なオーダによる confusion forest (SF) と、confusion network (CN)、システム出力の最大最小の BLEU 値を表 2 に示す*3。最大の BLEU と統計的に有意な差が見られなかった結果を太字で示す。confusion forest は $v = \infty, h = \infty$ としたとき、confusion network とほぼ同様な結果が得られた。これは、各システムの出力のルートからの導出を記憶し、その並びを保ったまま、木構造の断片を組み合わせることに相当する。 $v = \infty, h = 2$ の時、confusion forest は木単位の多少の並び替えが行われており、3つの言語対において最も良い結果がえられ、また cz、de では、confusion network と比較して統計的に有意な差が見られた。また、オーダを小さくすると、大幅な終端記号の並び替えが行われ、BLEU が小さくなることが確認された。これは、木構造を捉えた素性を導入することにより解決できると考えられ、今後の課題としたい。

表 3 にて平均ハイパーエッジ数で示されるハイパーグラフの大きさを示す。このように、confusion forest は confusion network と比べ、桁違いに小さいことがわかる。

5 おわりに

本稿は、confusion forest に基づく機械翻訳システムコンビネーションを提案した。単語の近さに基づいて複数の出力を confusion network で表現するのに対し、統語的な近さを利用した森である confusion forest で表現した。構文森は、システムの出力の構文解析木から局所的な文法を獲得することにより、生成される。実験結果から非常にコンパクトなデータ構造で confusion network とほぼ同様な結果を得られた。

表 2 BLEU による翻訳結果

言語対	cz-en	de-en	es-en	fr-en
システム最小	14.09	15.62	21.79	16.79
最大	23.44	24.10	29.97	29.17
CN	23.70	24.09	30.45	29.15
$CF_{v=\infty, h=\infty}$	24.13	24.18	30.41	29.57
$CF_{v=\infty, h=2}$	24.14	24.58	30.52	28.84
$CF_{v=\infty, h=1}$	24.01	23.91	30.46	29.32
$CF_{v=5, h=\infty}$	23.93	23.57	29.88	28.71
$CF_{v=5, h=2}$	23.82	22.68	29.92	28.83
$CF_{v=5, h=1}$	23.77	21.42	30.10	28.32
$CF_{v=4, h=\infty}$	23.38	23.34	29.81	27.34
$CF_{v=4, h=2}$	23.30	23.95	30.02	28.19
$CF_{v=4, h=1}$	23.23	21.43	29.27	26.53

表 3 ハイパーエッジの平均数 ($h = 1$)

言語対	cz-en	de-en	es-en	fr-en
CN	2,222.68	47,231.20	2,932.24	11,969.40
lattice	1,723.91	41,403.90	2,330.04	10,119.10
$CF_{v=\infty}$	230.08	540.03	262.30	386.79
$CF_{v=5}$	254.45	651.10	302.01	477.51
$CF_{v=4}$	286.01	802.79	349.21	575.17

これは、複数のシステムの出力を構文森として表現することにより、統語論的な同意を直接取る、初めてのシステムコンビネーション手法である。この confusion forest に基づくシステムコンビネーション手法には、将来発展の余地があると考えられる。例えば、統語論的な素性は、大きな森から仮説を識別するのに役立つと思われるが、本稿における 4.2 では用いていない。将来、解析エラーと confusion forest の生成との何らかのトレードオフ関係が存在するのか、解析器の精度を調整することにより分析を行いたい。文法に基づき森を生成したが、これに対して、木のノードに対する挿入、削除、置換を計算する木編集距離 (tree edit distance) [31] に基づく木アライメントの手法を試したい。

*3 例えば、 $h = 1$ の場合、左右のコンテキストを見て、最大 3 つのラベルが結合される。

参考文献

- 1 J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," In Proc. of ASRU, pp. 347–354, Dec. 1997.
- 2 Lidia Mangu, Eric Brill, and Andreas Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," *Computer Speech & Language*, 14(4): 373–400, 2000.
- 3 John C. Henderson and Eric Brill, "Exploiting diversity in natural language processing: Combining parsers," In IN PROCEEDINGS OF THE FOURTH CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, pp. 187–194, 1999.
- 4 Srinivas Bangalore, German Bordel, and Giuseppe Riccardi, "Computing consensus translation from multiple machine translation systems," In Proc. of ASRU, pp. 351–354, 2001.
- 5 Evgeny Matusov, Nicola Ueffing, and Hermann Ney, "Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment," In Proc. of EACL, pp. 33–40, 2006.
- 6 Xiaodong He, Mei Yang, Jianfeng Gao, Patrick Nguyen, and Robert Moore, "Indirect-HMM-based hypothesis alignment for combining outputs from machine translation systems," In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pp. 98–107, Honolulu, Hawaii, Oct. 2008. Association for Computational Linguistics.
- 7 Shyamsundar Jayaraman and Alon Lavie, "Multi-engine machine translation guided by explicit word matching," In Proceedings of the ACL 2005 on Interactive poster and demonstration sessions, ACL '05, pp. 101–104, Morristown, NJ, USA, 2005. Association for Computational Linguistics.
- 8 K. C. Sim, W. J. Byrne, M. J. F. Gales, H. Sahbi, and P. C. Woodland, "Consensus network decoding for statistical machine translation system combination," In Proc. of ICASSP, Vol. 4, pp. IV-105–IV-108, April 2007.
- 9 Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul, "A study of translation edit rate with targeted human annotation," In Proc. of AMTA, pp. 223–231, 2006.
- 10 Sylvie Billott and Bernard Lang, "The structure of shared forests in ambiguous parsing," In Proc. of ACL, pp. 143–151, June 1989.
- 11 Haitao Mi, Liang Huang, and Qun Liu, "Forest-based translation," In Proceedings of ACL-08: HLT, pp. 192–199, June 2008.
- 12 Jay Earley, "An efficient context-free parsing algorithm," *Communications of the Association for Computing Machinery*, 13: 94–102, Feb. 1970.
- 13 Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan, "Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation," In Proc. of WMT, pp. 17–53, July 2010.
- 14 Franz Josef Och and Hermann Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, 29(1): 19–51, 2003.
- 15 Antti-Veikko Rosti, Bing Zhang, Spyros Matsoukas, and Richard Schwartz, "Incremental hypothesis alignment for building confusion networks with application to machine translation system combination," In Proc. of WMT, pp. 183–186, June 2008.
- 16 Antti-Veikko Rosti, Spyros Matsoukas, and Richard Schwartz, "Improved word-level system combination for machine translation," In Proc. of ACL, pp. 312–319, June 2007.
- 17 Dan Klein and Christopher D. Manning, "Parsing and hypergraphs," In Proc. of IWPT, pp. 123–134, 2001.
- 18 Stuart M. Shieber, Yves Schabes, and Fernando C. N. Pereira, "Principles and implementation of deductive parsing," *Journal of Logic Programming*, 24(1-2): 3–36, July-Aug. 1995.
- 19 Dan Klein and Christopher D. Manning, "Accurate unlexicalized parsing," In Proc. of ACL, pp. 423–430, July 2003.

- 20 Joshua Goodman, "Semiring parsing," Computational Linguistics, 25: 573–605, Dec. 1999.
- 21 Liang Huang and David Chiang, "Better k-best parsing," In Proc. of IWPT, pp. 53–64, Oct. 2005.
- 22 David Chiang, "Hierarchical phrase-based translation," Computational Linguistics, 33(2): 201–228, 2007.
- 23 Liang Huang and David Chiang, "Forest rescoring: Faster decoding with integrated language models," In Proc. of ACL, pp. 144–151, June 2007.
- 24 Adam Lopez, "Translation as weighted deduction," In Proc. of EMNLP, pp. 532–540, March 2009.
- 25 Shankar Kumar, Wolfgang Macherey, Chris Dyer, and Franz Och, "Efficient minimum error rate training and minimum bayes-risk decoding for translation hypergraphs and lattices," In Proc. of ACL/IJCNLP, pp. 163–171, Aug. 2009.
- 26 Antti-Veikko Rosti, Necip Fazil Ayan, Bing Xiang, Spyros Matsoukas, Richard Schwartz, and Bonnie Dorr, "Combining outputs from multiple machine translation systems," In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, pp. 228–235, Rochester, New York, April 2007. Association for Computational Linguistics.
- 27 Wolfgang Macherey and Franz J. Och, "An empirical study on computing consensus translations from multiple machine translation systems," In Proc. of EMNLP-CoNLL, pp. 986–995, June 2007.
- 28 Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," In Proc. of ACL, pp. 311–318, July 2002.
- 29 Zhifei Li and Sanjeev Khudanpur, "Efficient extraction of oracle-best translations from hypergraphs," In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, pp. 9–12, Boulder, Colorado, June 2009. Association for Computational Linguistics.
- 30 Markus Dreyer, Keith Hall, and Sanjeev Khudanpur, "Comparing reordering constraints for smt using efficient bleu oracle computation," In Proceedings of SSST, NAACL-HLT 2007/AMTA Workshop on Syntax and Structure in Statistical Translation, pp. 103–110, Rochester, New York, April 2007. Association for Computational Linguistics.
- 31 Philip Bille, "A survey on tree edit distance and related problems," Theor. Comput. Sci., 337: 217–239, June 2005.

(平成 24 年 6 月 14 日 採録)



わたなべ たろう
渡辺太郎

ユニバーサルコミュニケーション研究所
多言語翻訳研究室主任研究員
博士 (情報学)
機械学習、機械翻訳、自然言語処理
taro.watanabe@nict.go.jp