単語類似度の尺度比較可視化ツールの作成

河部恒 †‡　柏岡秀紀 †‡　田中英輝 §　松本裕治 †

† 奈良先端科学技術大学院大学 情報科学研究科
‡ ATR 音声コミュニケーション研究所
§ ＮＨＫ放送技術研究所
{kou-k,matsu}@is.aist-nara.ac.jp,
hideki.kashioka@atr.co.jp,
tanaka.h-ja@nhk.or.jp

1　はじめに

単語は、自然言語における処理の単位としてもっとも基本的なものであり、特に単語をベースとした類似度の比較は様々な分野で欠かすことの出来ない要素技術となっている。単語の表す内容を比較する上で、各単語がどのくらい似ているのかを表す尺度として、適当な距離空間を導入して単語間の類似度を定義するという方法が一般的であるが、パラメータの細かな調整などを行う場合、この変更が結果にどう反映しているかが把握しにくい点が問題であった。今回我々が作成した単語類似度の可視化システムでは、こうした変更の直感的な理解を可能にした。またアプリケーションとして用例翻訳システムにおける入力文と用例文の類似度の比較を取り上げ、KL-divergence, Information radius, alpha-skew, L1 norm などの尺度を用いて計算し、それらを比較し検討を行ったのでそれを報告する。

2　距離行列の作成

単語間の類似度を計算するモジュールはコーパスおよびシソーラスを入力とし、類似度を計算したい単語リストおよび、類似度尺度を与えると距離行列を作成するようになっている。(図 1)
詳細は [6] を参照のこと。

以下に確率モデルベースの類似度尺度の例を示す。

$$IR(p,q) = \frac{1}{2}\left[D\left(p \parallel \frac{p+q}{2}\right) + D\left(q \parallel \frac{p+q}{2}\right)\right]$$

$$\alpha\text{-}skew\,(p,q) = D\left(\, q \parallel \alpha \cdot p + (1-\alpha)\cdot q\right)$$

$$L_1(p,q) = \sum_i \mid p_i - q_i \mid$$

[1][4][2][5][3]

3　可視化システム概要

可視化システムは JavaSDK 1.4.2-01 で開発し、Windows 2000,XP 上で動作確認をしている。

画面は三つのペインに分かれる。すなわちパラメータ表示部、メインウィンドウ、検索単語入力部である。(図 3)

[6] によりコーパスおよびシソーラスから $N * M$ の距離行列がつくられる。(但し N = 対象単語 (数百程度)、M =共起単語 (数千程度))

この距離行列をもとに、単語はボトムアップにクラスタリングされ一本のツリーになる。ツリーのリーフには各単語が現れ、類似度に応じてアークの長さが変化する。なお最終的なツリーは XML 形式で保存される。

可視化を容易にするため、ツリー全体は (1) 拡大/縮小/回転、(2) 平行移動、(3) 特定単語からのノード段数の展開数の調整、を全てスクロールバーで行えるようになっており、また各単語はマウスでドラッグすることにより位置を移動することができ、直感的な視覚化が可能になっている。

検索単語入力部に単語を入力すると、類似度の小さい順に右側部分に単語リストと類似度が表示され、同時にメインウィンドウに当該単語とその周辺数単語が展開される。ツリーは巨大になることがあるため、はじめから全ては展開されない。

4　データ

今回用いたコーパスおよびシソーラスは以下の通りである。

- コーパス: NHK News Corpus 1995-2000
- シソーラス: 角川類語新辞典 [7]

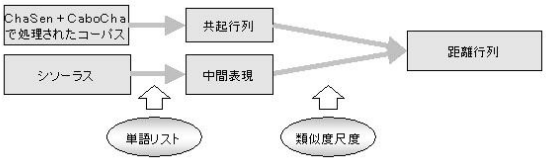

図 1: システムの構成

図 2: システムの構成

デモでは類似度尺度の種類、コーパスの量、共起を取るときに window size, 共起の種類などによって類似度の構造全体がどう変化するかを紹介する。

5 謝辞

本研究は通信・放送機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである。

参考文献

[1] Robert Dale, Hermann Moisl, and Harold Somers, editors. *Handbook of Natural Language Processing, Chap.19*. Marcel Dekker, 2000.

[2] Lillian Lee. Measures of distributional similarity. In *37th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (ACL) *University of Maryland, USA*, pp. 25–32, 1999.

[3] Dekang Lin. An information-theoretic definition of similarity. In *Machine Learning *Proceedings of the Fifteenth International Conference (ICML '98)*, pp. 296–304, 1998.

[4] Manning and Schuetze. *Foundations of Statistical Natural Language Processing, Chap.8.* The MIT Press, 1999.

[5] Eiichiro Sumita and Hitoshi Iida. Experiments and prospects of example-based machine translation. In *29th Annual Meeting of the Association for Computational Linguistics * Proceedings of the Conference*, pp. 185–192, 1991.

[6] 河部恒, 柏岡秀紀, 田中英輝, 松本裕治. 単語類似度の尺度比較支援ツールの作成. 情報処理学会自然言語処理研究報告 NL-156-06, 2003.

[7] 大野晋, 浜西正人. 角川類語新辞典. 角川書店, 1981.

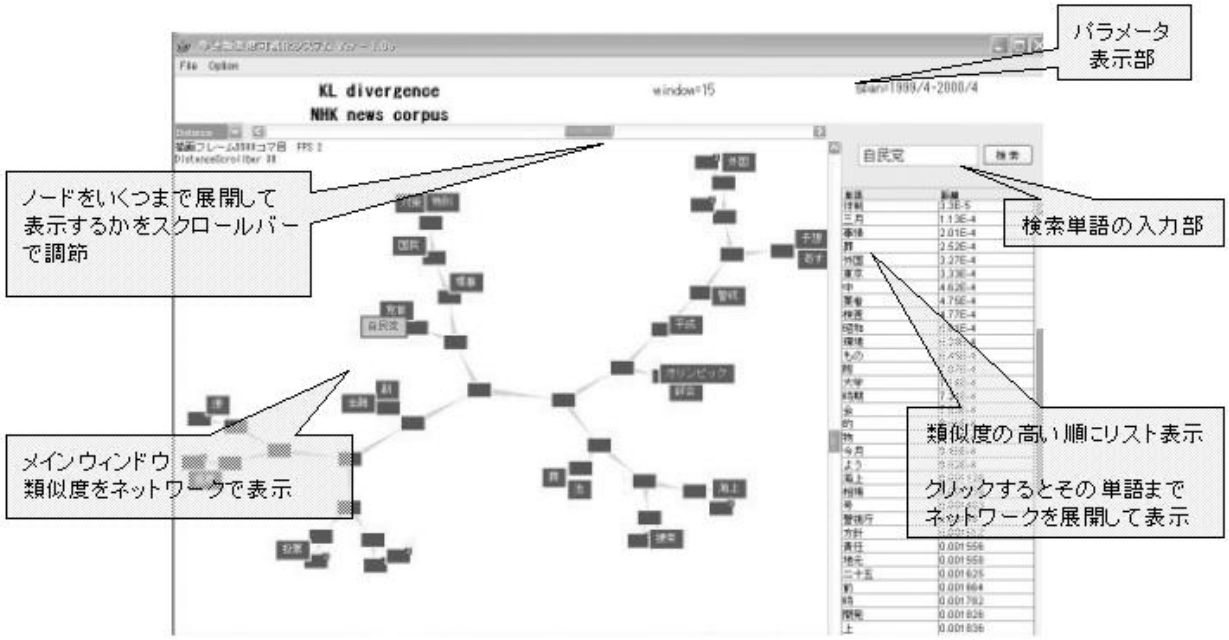

図 3: システム画面