

知能工学部門 認知機構研究分野 (奥村学研究室)

Email : oku@pi.titech.ac.jp

http://www.lr.pi.titech.ac.jp/

(研究分野)

奥村研究室では、自然言語処理とその応用分野を研究しています。

(研究テーマ)

1) テキストの局所的一貫性モデル (横野光)

テキストの評価は、自動要約などのようなテキストを生成するタスクにおける手法の評価として用いられるだけでなく、小論文の自動評価などの教育的な用途にも用いられる。テキストの評価には、内容の正しさに関する評価と、読みやすさや一貫性などのテキストの構造に関する評価が考えられる。本研究では、構造に関する評価に関して、隣接する文間の意味的なつながりである局所的一貫性に焦点を当てる。テキストの文間の文法的なつながりである結束性が局所的一貫性と互いに関係しているという仮定に基づき、接続関係や語彙的結束性などの結束性に関わる要素を組み入れた局所的一貫性モデルを提案する(図1, 4)。

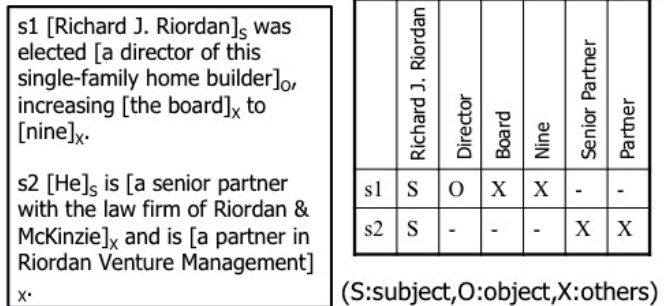
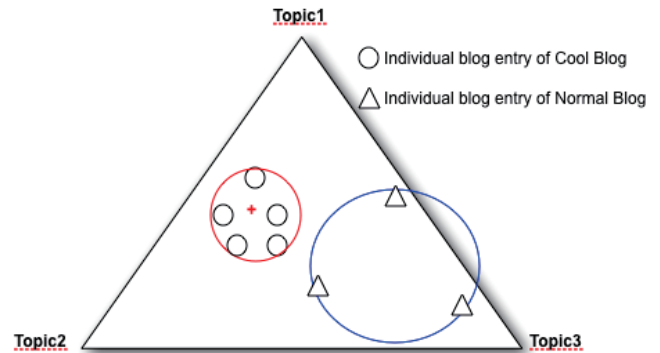


図1 テキストの例とそのentity grid表現

Fig.1 Example of text and its entity grid representation

2) Cool Blog Classification (Kritsada Sriphaew)

インターネット上に存在する無数の中には、役に立つブログがある一方、読者にとって有益とはいえないものも多くある。読者に有益なブログを提供するため、本研究では、面白いブログを自動的に発見する技術を開発する。面白さは主観的な概念ではあるが、専門性が高くまた内容に一貫性のあるブログは読者の興味を惹きやすいであろう。そのようなブログを同定するため、ブログにおける様々な話題の生起確率やその一貫性、あるいはエン트리数などの情報を利用する技術を開発した。図2は、話題の生起確率を表すシンプレックスであり、面白いブログがどのように分布するかを例示している。また、図5は面白いブログのデータとラベル無しデータからサンプリング技術を用いて分類器を構築する方法を示している。



Simplex of Topics where each item in the simplex represents the topic probability

図2 面白いブログのトピック確率の特性

Fig.2 Characteristics of topic probability for cool blogs

3) WEB上の画像の分類とメタデータ付与による携帯電話向けWEB表示 (成川夏子)

近年、回線の高速化やパソコンの大画面化に伴いWEB上の画像が増えている。一方で、携帯電話からWEBを利用する機会も増えている。しかし、既存の携帯電話向けのブラウザでは画像に対して特別な処理を行っておらず、そのまま画像を表示したり縮小したりしているのみである。そこで本研究では、WEBページからの言語処理的情報と画像そのものの画像処理的情報を組み合わせて利用することにより、WEB上の画像を性質、構造、重要度の3つの基準に基づいて分類し(図3)、さらにそれぞれの画像に内容を表すメタデータを付与し、それらの情報を利用して画像の最適な表示方法を類推することで、より使いやすい携帯電話向けのWEB表示を作成する(図6)。

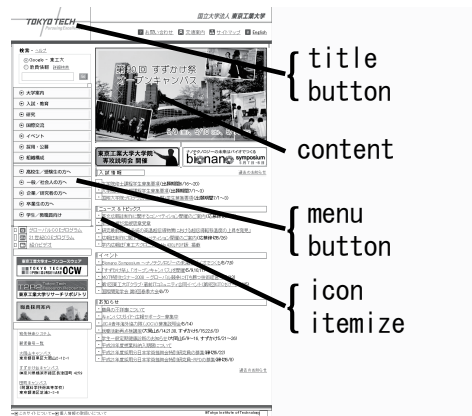


図3 WEB上の画像の分類

Fig.3 Classification of images on webpages

Advanced Information Processing Division

Information Processing and Recognition Section

(Manabu Okumura Group)

(Research Field)

Natural language processing and its applications

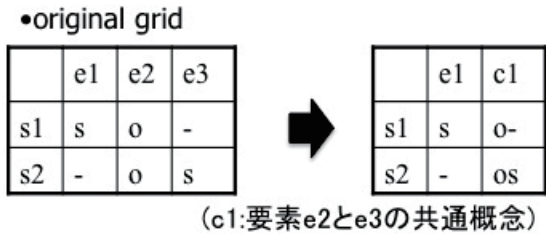
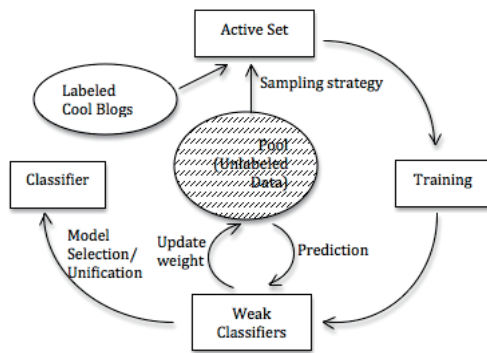


図4 語彙的結束性に基づく要素のクラスタリング

Fig.4 Entity clustering based on lexical cohesion



Bagging technique to learn a classifier from unlabeled data

図5 正例とラベル無データからの分類器構築

Fig.5 Classifier construction with positive and

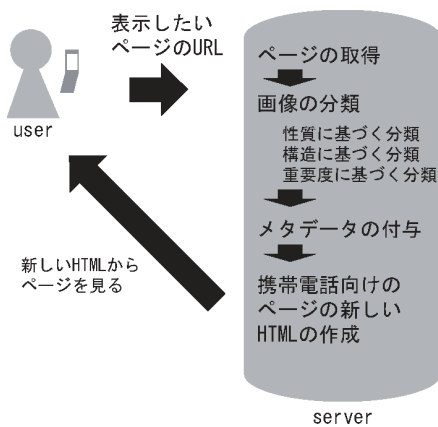


図6 携帯電話向けWEB表示のプロセス

Fig.6 Mechanism of cellphone browsing of webpages

(Current Topics)

1) Local coherence of text (Hikaru Yokono)

Models of text evaluation have been often used for evaluating not only humanwritten text in educational applications but also automatically created text in text generation, such as automatic summarization. There are two aspects for evaluating text: adequacy of content and quality. Text quality tends to be assessed with respect to grammaticality, coherence, etc. We concentrate on modeling and evaluating text coherence, which is the semantic relationship between adjacent sentences. We propose local coherence model incorporating cohesive devices such as conjunctions and lexical cohesion, assuming that text cohesion is related to the coherence (Figs. 1 and 4).

2) Cool Blog Classification (Kritsada Sriphaew)

A massive number of blogs has led to not only huge amount of informative content, but also abundance of uninteresting content. To help in filtering useful content, this research focuses to identify the *cool* blogs from numerous blogs. We believe that the blogs with high expertise knowledge and consistency are likely to draw attention of the readers. The characteristic of cool blog against normal blog can be demonstrated on topic simplex diagram in Fig. 2. Fig. 5 depicts the procedure of classification when we have only cool blogs and unlabeled blogs for training a classifier.

3) Image classification and metadata creation for mobile WEB browsing (Natsuko Narikawa)

Today we often use mobile phones for browsing webpages with many images. Most mobile phones, however, do not have a browser that works well under a limited bandwidth. We classify images according to their properties, structures, and importance (Fig. 3), so that we can use the classification result for deciding whether we should display, shrink, or delete each image. Finally, we propose a new WEB browsing system for mobile phones which uses the classification result and the metadata to render webpages with better readability (Fig. 6).