観光ガイドシステムに必要な知識のWeb文書からの自動獲得

柿澤 康範 (710017)
北陸先端科学技術大学院大学 情報科学研究科

2009年2月5日

キーワード: 属性情報，トラブル，Web文書，大規模コーパス.

本論文は，Web文書から自動獲得した知識を選別することで，ユーザが必要とする情報を無駄無く提供できるような観光ガイドシステムを構築することを目指している．Web文書から自動獲得した知識として，本論文では属性情報とトラブル情報を扱う．

ここで属性とは，人が知りたい対象物の側面（例えば寺であれば,「拝観するのにかかる料金」や「寺に行くための方法」）のことであり，文書中では具体的な**属性語**（例：「拝観料」,「交通手段」）によって参照される．そして属性語で実際に示される値を**属性値**（例：属性語「拝観料」→属性値「300円」）と呼ぶ．この属性語／属性値のことを**属性情報**と呼ぶ．属性情報は，吉永らによってWeb文書集合からの自動獲得 [1] が行われた．また，トラブル情報についても，De Seagerらによって自動獲得 [2] が行われた.（例：ディズニーランドのトラブル →「身長制限」,「渋滞」）

このようなWeb文書から自動獲得された知識は，ユーザに情報提供を行うシステムの知識源として利用できる．しかし，単純に対象物に属する情報を全て提示してしまうと，情報量が多すぎるため，ユーザ自身で情報を選別する必要がある．そこで，ユーザにとって必要な情報のみを選別して提供することで，一度に示される情報を減らし，ユーザ自身で情報を選別しなくても済むようにしたい．

このような知識の選別を行うため，以下の３つの知識獲得を行った．

ユーザがとる行為を表す動詞による属性情報の分類　属性情報を，ユーザがとる行為を表す動詞（「行く」や「見る」など）で分類した．これにより，ユーザが「清水寺に行く」といったときに「交通手段」や「住所」を示し,「清水寺を見たい」といったときに「見所」を示すことができる．

分類手法として，属性語と係り受け関係にある頻度が最も大きい動詞を分類結果とするベースラインと，”＜名詞＞の＜属性語＞”というパターンに当てはまる名詞と係り受け関係にある頻度が最も大きい動詞を分類結果とする提案手法を評価した．その結果，ベースラインが26%，提案手法が42%であり，提案手法による精度の向上を確認した．ただし，まだ精度が低いので改善の余地がある．

トラブル動詞によるトラブル名詞の分類　「渋滞」,「人混み」といったトラブルを表す名詞（以下，**トラブル名詞**）を,「遅れる」,「疲れる」といったトラブルによって引き起こされる事象を表す動詞（以下，**トラブル動詞**）で分類する．これにより，そのトラブルがどのような問題を起こすのかがわかる．例えば,「白飛び」,「こむら返り」といったトラブルを知らないユーザがいるとして,「白飛びで撮れない」,「こむら返りで痛む」という形でトラブル動詞も同時に示すことで，そのユーザはトラブルの意味を大まかに知ることができる．

分類手法として，”＜トラブル名詞＞で＜トラブル動詞＞”というパターンの頻度が最も大きいトラブル動詞を分類結果とするベースラインと，ベースラインのスコアを，トラブル名詞とトラブル動詞の深刻度の差が大きいほどスコアが小さくなるようにした提案手法を評価した．その結果，提案手法ではベースラインから改善が見られなかった．なお，ベースラインの精度は 84%だった．

トラブル動詞の深刻度のランク付け　トラブル名詞の深刻度が分かれば，ユーザに深刻度の大きいトラブルを選別して示すことができる．例えば,「海水浴場」のトラブル情報に「水難事故」と「日焼け」があったとき，深刻度の大きい「水難事故」を優先して提示することができる．ただし，必ずしも深刻度の大きいトラブルを提示すれば良いとは限らない．この点は今後検討する余地がある．

トラブル名詞の深刻度を求めるために，トラブル動詞の深刻度のランク付けを行った．トラブル動詞の深刻度のランク付けができれば，トラブル動詞の深刻度の値も大まかに得ることができ，トラブル名詞の分類先のトラブル動詞の深刻度を見ることで，トラブル名詞の深刻度も推定できる．ランク付けの方法として，シェッフェの一対比較による５分類を，学習データに対して人手で行い，機械学習によって残りのデータも一対比較のデータを得た．機械学習法としてはSVM と最大エントロピー法を評価した．その結果，SVM は 65%，最大エントロピー法は 68%の精度となった．(特定の条件での２分類では 97%)

今後の課題　今後の課題としては，提案手法の改善と，ユーザの行動プランの知識の自動獲得を行う．また来年度には，本研究で獲得した属性情報とトラブル情報に関する知識を，実世界の音声対話システムに組み込む計画を立てている．

参考文献

[1] Naoki Yoshinaga and Kentaro Torisawa, ”Open-Domain Attribute-Value Acquisition from Semi-Structured Texts” In Proceedings of the Workshop on Ontolex 2007 – The Lexicon/Ontology Interface held at the sixth International Semantic Web Conference, pp. 55-66. Nov., 2007.

[2] S. De Saeger, K. Torisawa, and J. Kazama. Looking for trouble. In Proc. of The 22nd International Conference on Computational Linguistics (Coling2008), 2008.