Semantic Lexicon Adaptation for Use in Query Interpretation

Ana-Maria Popescu
Yahoo! Labs
Sunnyvale, CA
94089
amp@yahoo-inc.com

Patrick Pantel
Yahoo! Labs
Sunnyvale, CA
94089
ppantel@yahoo-inc.com

Gilad Mishne
Yahoo! Labs
Sunnyvale, CA
94089
gilad@yahoo-inc.com

ABSTRACT

We describe improvements to the use of *semantic lexicons* by a state-of-the-art query interpretation system powering a major search engine. We successfully compute *concept label importance* information for lexicon strings; *lexicon augmentation* with such information leads to a 6.4% precision increase on affected queries with no query coverage loss. Finally, *lexicon filtering* based on label importance leads to a 13% precision increase, but at the expense of query coverage.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Storage and Retrieval

General Terms

Algorithms, Experimentation, Measurement

1. INTRODUCTION

An important aspect of query interpretation is recognizing the *types* of entities mentioned in a user query. We focus on a particular aspect of entity tagging in queries: the use of semantic domain knowledge in the form of *semantic lexicons*, whose limitations can include *limited coverage*, *noise* and finally, *ambiguity* of lexicon entries, which is the focus of this poster. A high-precision, large semantic lexicon extracted from Web data may contain few ontological errors, but many ambiguous strings - some with a different primary sense: e.g., given an Actor lexicon, `clint eastwood` and `mel gibson` are strings for which Actor is an *important* sense, while `snoop dogg` is not. Our work computes *label importance* information for lexicon strings in order to enrich the semantic lexicons supplied to a client query interpretation system for better performance.

Web Query Interpretation with QIS

QIS is a state-of-the-art query interpretation engine which powers a major search engine. Given a query, QIS produces the most likely *interpretations* for it: an interpretation consists of a set of a semantic types assigned to non-overlapping query tokens. QIS makes use of a large type taxonomy and uses a combination of taggers (both statistical and semantic lexicon lookups) in order to generate candidate *interpretations* of the query. Next, a ranking module assigns a score to each interpretation; ranking is approached as a supervised learning task, where manual judgments of the quality of interpretations for randomly-selected queries are used as training data. A multitude of features is used to represent each interpretation, including both *query-level features* (e.g., number of words, the name of taggers which participated in forming the interpretation) and *tag-level features* (e.g., tagger confidence).

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

Table 1: Semantic Lexicon Entries: Examples

Actor = Important label	**Actor = Minor label**	**Actor = Incorrect label**
al pacino	karen mulder	girls
robert de niro	sharon osborne	han solo
robert redford	snoop dogg	mark twain
tom hanks	martin scorsese	hollywood

2. CONCEPT LABEL IMPORTANCE

We describe an *annotation* task which estimates *concept label importance* for strings in a lexicon and is used for gold standard data set construction.

Importance Annotation Consider the string s = `al pacino` in the Actor lexicon. We retrieve the queries containing s and rank them by their frequency; we retain the top k queries whose cumulative frequency is more than 80% of the *total* frequency of queries containing s. For each query q, we manually examine the top 10 web search results and record the % of pages r in which `al pacino`'s Actor sense appears dominant. The score for (Actor, s) is the sum of r_1, ... r_k, each weighted by the relative query frequency for q_i: e.g., the scores for the *Actor* label and `al pacino`, `newman` are 1 and, respectively, 0.78. We use 0.2 as a manually chosen *importance threshold* on score values.

We detect label importance using two main methods:

M1. I_{CM} uses a *supervised ML* approach (Gradient Boosted Decision Trees [1]). The *training set* is a sample of the target lexicon manually labeled as described above. *Features* include generic string-level ambiguity features and features for concepts in the QIS taxonomy (see Table 2).

M2. Given a set of concepts and a string s, I_{MM} computes an estimate of $P(C_i|s)$ for each concept C_i using the n-ary classification mode of GBDT and normalizing the resulting scores. 0.2 is used as the importance threhold on the resulting scores. The *training set* is computed automatically: positive examples for each class C are sampled from a pre-computed global cluster aligned with the target class

Table 2: Features for lexicon element e and concept C

Description	Source
capitalization ratio, string freq.	Web corpus
relative freq. of C in the set of tags for e's pages	Wikipedia
distrib. sim. with centroid of aligned cluster CL % of centroid features shared with e distrib. sim. with centroid of S (concept seed set) % of centroid features shared with e, etc.	Web corpus
% of high PMI query/doc. contexts for S which cooccur with e, etc.	query log, clicked docs

Table 3: Set of experimental lexicons

Lexicon	Actor	Manufacturer	Movie	TV Show	Magazine Newspaper
Size	376,653	51596	120961	14012	5806

Table 4: Detecting concept label importance (F-1 measure)

Method	Actor	Man.	Movie	TVShow	Mag.
B_i	0.46	0.66	0.51	0.86	0.60
I_{CM}	**0.91**†	**0.87**†	0.74†	**0.87**	**0.88**†
I_{MM}	0.84†	0.60	**0.87**†	0.61	0.78†

(see Feature Set discussion below).

Feature Set In addition to string ambiguity clues and Wikipedia category information, we derive a large set of *data source-specific concept representations* and features from a large 2008 Web crawl, 1 year worth of query logs and 3 months' worth of queries together with clicked Web pages (examples in Table 2). *Local* concept representations are a) the centroid of a concept seed set S or b) the top 100 *sentence*, *query* or *document contexts* ranked by average PMI (Pointwise Mutual Information) with the elements of S. *Global* representations are 2597 semantic clusters computed using the Cluster-by-Committee algorithm [2] on 1/5th of the 2008 Web crawl and automatically correlated with the lexicons and concepts of interest based on the relative overlap in the element set or centroid feature set.

Detecting Label Importance: Evaluation

For each of the 5 lexicons in Table 3 , we randomly sample 400 examples from the intersection of each semantic lexicon with a set of frequent queries and annotate them as described above. We compare I_{CM} and I_{MM} with a baseline B_i which assumes that the target label is always *important* for a given lexicon string (a 10-fold cross validation setup is used). I_{MM} uses a separate (automatically constructed) training set, but is evaluated on the same test set subsets as I_{CM}. The results in Table 4 show that *concept label importance can be reliably detected across lexicons*, especially by I_{CM} (improvements over baseline are statistically significant at the 0.95 level). Errors are due to the lexical diversity of media lexicons (e.g., "omen", "gandhi", "night" are all Movies), noisy global clusters (`lauren london` in the Musician cluster), etc.

2.1 QIS with Enhanced Lexicons

Given the 5 target lexicons, we experiment with the following lexicon modification types:

1. **Lexicon Filtering** $f(I_{CM})$ and $f(I_{MM})$ denote two filtering (string removal) operations based on the concept label importance information computed by I_{CM} and I_{MM}. I_{MM} is also used to derive a *dominance* filter $f(D_{MM})$, which eliminates a string s from lexicon L for concept C is C is not the most likely label for s.

KB Version	P@1 *gen aff*	Cov *gen aff*	P@1 *sig aff*	Cov *sig aff*	P@1 *test set*	Cov *test set*
original	74.7	546	69.1	307	81.5	1027
$a(I_{CM})$	81.1†	546	78†	307	84.8†	1027
$a(I_{MM})$	80.3†	546	75.1†	307	82.9	1027
$f(I_{CM})$	**87.8**†	288	**89.7**†	247	88.4†	1038
$f(I_{MM})$	84.8†	302	87.2†	253	86.2†	1034
$f(D_{MM})$	86.7†	271	88.3†	234	**89.3**†	1035

Table 5: Positive Impact of Semantic Enhancements on Lexicon Use in Query Interpretation

2. **Lexicon Augmentation** $a(I_{CM})$ and $a(I_{MM})$ denote feature additions based on *non-binary* label importance scores computed by the *regression* versions of I_{CM} and I_{MM}: given an interpretation I of a query q which contains tokens t_1, ... t_k of type C, we use the importance scores $imp(t_i, C)$ to derive interpretation-level features: e.g., $f(I)_{Actor(avg)} = \sum_1^k imp(Actor, t_i)/k$.

The research version of QIS is trained on 27,000 queries and separately tested on 2700 queries. Evaluating performance amounts to evaluating the *ranking* of query interpretations; the gold standard is represented by manually judged interpretations for a randomly selected set of queries. *Relevant* queries have at least 2 interpretations, at least one of which is judged as *Excellent* or *Good* in the manually labeled data. *Test set* queries are relevant queries for which interpretations have been generated based on the *comprehensive* set of QIS lexicons. *Generally affected* queries are relevant queries with at least 1 interpretation containing a semantic tag which matches a target concept; for *Significantly affected* queries, this tag belongs to a *Good* or *Excellent* interpretation.

Metrics and Discussion *Precision*@1 is the precision for the top ranked interpretation; *Coverage* is the number of relevant queries for which interpretations are generated. As seen in Table 5, adding features based on label importance information leads to an increase in precision on affected queries of 6.4% and an overall precision increase of 3.3% (both significant at the 0.95 level). Lexicon filtering based on label importance leads to significant precision increases: 13% on affected queries and 7% overall, but at the cost of a 23% reduction in queries significantly affected and 48% in the number of queries generally affected.

Related Work [3] extracts lexicons from structured Web data and filters them for NER use in product queries, with a reported 25% increase in word-level tagging performance over a lexicon-free baseline. We show that *augmenting* lexicons is a promising alternative to filtering and report on *query interpretation* performance in *generic* Web search rather than *word-level tagging* performance in a particular domain.

3. REFERENCES

[1] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[2] P. Pantel and D. Lin. Discovering word senses from text. In *Proceedings of KDD*, pages 613–619, 2002.

[3] Y. Wang, R. Hoffmann, X. Li, and J. Syzmanski. Semi-supervised learning of semantic classes for query understanding - from the Web and for the Web. In *Proceedings of the CIKM*, 2009.