

Automatically Annotating Textual Resources with Human Intentions

Markus Strohmaier

Graz University of Technology and
Know-Center
Inffeldgasse 21a
8010 Graz, AUSTRIA

markus.strohmaier@tugraz.at

Mark Kröll

Graz University of Technology
Inffeldgasse 21a
8010 Graz, AUSTRIA

mkroell@tugraz.at

Christian Körner

Graz University of Technology
Inffeldgasse 21a
8010 Graz, AUSTRIA

christian.koerner@tugraz.at

ABSTRACT

Annotations represent an increasingly popular means for organizing, categorizing and finding resources on the “social” web. Yet, only a small portion of the total resources available on the web are annotated. In this paper, we describe a prototype - iTAG - for automatically annotating textual resources with human intent, a novel dimension of tagging. We investigate the extent to which the automatic analysis of human intentions in textual resources is feasible. To address this question, we present selected evidence from a study aiming to automatically annotate intent in a simplified setting, that is transcripts of speeches given by US presidential candidates in 2008.

Categories and Subject Descriptors

H.1.2 [User/Machine Systems]: Human Factors, H.3.1 [Content Analysis and Indexing]: Linguistic Processing, H.4.m [Information System Applications]: Miscellaneous

General Terms

Algorithms, Experimentation, Human Factors.

Keywords

Automatic Tag Generation, Human Intentions, Folksonomies

1. INTENT ANNOTATION

Existing automatic tag suggestion approaches largely focus on annotating a document according to its predominant subject matter (what a resource is about [1], e.g. “sports” or “politics”). In this work, we annotate resources according to the intentions described within them (what *goals* a resource is about, e.g. “Achieve Happiness” or “Maintain Good Health”). In contrast to topic, sentiment or opinion tags, *Intent Annotations* deal with *future* states of affairs that some agent wants to achieve (goals). Using a sample of web documents we observed that people rarely state their intentions explicitly in text, which makes the task of *Intent Annotation* especially challenging. As an example, consider the human intent category “Achieve Salvation” (taken from [1]). Although this is an activity pursued by many, it is extremely rare to find someone who states their plan on how to accomplish this goal. However, people are quite prolific in writing about the actions and activities they participate in on a

daily basis, such as “*adhere to Jewish law*” or “*convert to Christianity*”, which can be assumed to indirectly contribute to “*Achieve Salvation*”.

In this research, we explore the use of such *indicative actions* as a proxy for inferring intentions from textual resources. Intent Annotation can be understood as the problem of identifying a set of adequate intent annotations for each and every *action* indicative of intent in a given textual resource. More formally: Let $A^I = \{a^1_1, a^1_2 \dots a^1_n\}$ be a set of intent annotations and R be a domain of resources. Each textual resource r_i comprises a sequence of sentences $S = \{s_1, s_2 \dots s_{|S|}\}$. The task of Intent Annotation is to approximate the unknown function $f: S \times A^I \rightarrow \{True, False\}$, thereby assigning intent annotations ranging from 0 to $|A^I|$.

2. AUTOMATIC APPROACH

The iTAG approach presented in this paper consists of the following building blocks: In a first step, we use the Web as a resource to build up a knowledge base that maps indicative actions to intent categories of an existing goal taxonomy. In a second step, we scan a target text for indicative actions and look them up in the knowledge base to produce intent annotations.

Enriching a Taxonomy of Human Goals: We employed the social-psychological theoretical framework [1] that organizes high-level goals of people into 135 categories of human intent including “A good marriage”, “Getting an education” and “Taking care of family”, which are further grouped into high-level categories such as ‘Family’, ‘Religion’ and ‘Money’. In order to semantically enrich these descriptions, we attempted to find descriptive phrases for each category. To give an example: Descriptive phrases for the category “Achieve Salvation” included “*to reach spiritual enlightenment*” or “*to get into heaven*”. The process of enriching the taxonomy with phrases was iterative. Dr. Stephen Read, one of the co-authors of [2], evaluated these mappings to help us better understand category distinctions.

Constructing the Knowledge Base: We built a large knowledge base consisting of actions that indicate relevance for one of 135 categories. We attempted to acquire *indicative actions* by searching for sentences on the web that contained both (i) one of the descriptive phrases for the category, and (ii) an action-based causal relation. To achieve that, we constructed a series of query strings by concatenating each descriptive phrase with each of the following causal relation phrases: “in order to” and “for the purpose of”. Then, exact phrase searches were issued to the web using Yahoo! BOSS for all constructed query strings. The textual content of the first 500 result pages was retrieved and sentence delimited. Sentences that contained the query phrases were stored in our knowledge base, an Apache Lucene index.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.



Figure 1: Visual Comparison between Intent- and Traditional Tag Clouds based on Speeches Given by B. Obama and J. McCain

Table 1 shows sample queries and retrieved sentences. The minimum number of sentences per category was 12 (Category: “Firm Values”), the maximum number was 7,323 (Category: “Helping Others”) and the average number was 1,249.

Table 1: Exemplary Queries for the Category “Looking Young” and Corresponding Retrieved Sentences.

Query string causal relation + descr. Phrase	Retrieved Sentences (Yahoo) indicative actions
“in order to <i>look young</i> ”	In order to <i>look young</i> and beautiful, you need to <u>take care of your skin</u> .
“in order to <i>avoid wrinkles</i> ”	You need to <u>moisturize inside and out</u> , in order to <i>avoid wrinkles</i> .

Matching Sentences to Intent Categories: To automatically generate intent annotations, we first segment a given document into a set of sentences. Then, each sentence is issued as a query to the knowledge base. Using the default Lucene similarity measure, we identified the most similar sentence in our knowledge base containing indicative actions. The intent category associated with the knowledge base entry is then assigned as the intent annotation for the analyzed sentence in a ‘Winner takes it all’ approach. Intent annotations for documents are produced by aggregating intent annotations of all sentences.

3. APPLICATION AND EVALUATION

To gauge the prospects of intent annotation, we applied our approach to a limited set of textual documents that we suspected to be particularly amenable for our purposes. We retrieved and preprocessed the textual resources of 44 transcripts of political speeches given in April and June 2008 by the two American presidential candidates, J. McCain and B. Obama. After sentence delimitation, every sentence was treated as a query for the knowledge base. The results are depicted in Figure 1. On the left hand-side, Intent Tag Clouds of J. McCain’s and B. Obama’s speeches highlight the goal categories that are most important to them according to iTAG. On the right, traditional tag clouds give an overview of the vocabulary most used by the two candidates. While J. McCain’s most dominant intent annotations are “Helping Others” and “Being better than others”, “Pursuing ideals” and “Helping Others” represent the highest-weighted annotations for B. Obama. We also conducted several human subject studies to evaluate (i) the quality of the knowledge base and (ii) the quality of iTAG annotations. In the first study, a random sample of 674 knowledge base entries was evaluated according to the plausibility of the action/category mapping, with 57% of the entries containing indicative actions. In a second study, we had human subjects annotate 6399 sentences of Obama’s and McCain’s speeches. We used the top

25 annotations produced by our subjects as relevant annotations (our “ground truth”) and judged the remaining human annotations to be irrelevant. Figure 2 shows the performance of iTAG in comparison to a simple baseline approach that ranks intent categories in a random manner. The precision/recall curves show that iTAG outperforms the baseline for recall levels of up to 70%. While these results can certainly be improved, the overall outcome appears to be both interesting and plausible.

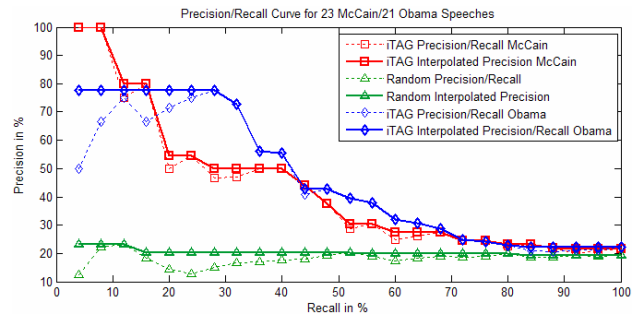


Figure 2: Evaluation of iTAG vs. a Random Approach

Like other dimensions of tagging, *intent* annotations add interesting information to textual resources, which is difficult to extract from the resource itself. Our work contributes a novel dimension to the set of tag dimensions identified in the literature. The iTAG prototype demonstrates the principle feasibility of automated intent annotation in a simplified setting (i.e. 44 political speeches) and thereby extends the repertoire of existing automatic tag generation techniques. In this sense, our work contributes to expanding the knowledge that can be inferred from textual resources. The iTAG approach could help to open up a new intentional dimension to navigating and browsing textual resources on the web.

Acknowledgements: Thanks to S. Read for assistance regarding the human goal taxonomy. Special thanks to A. S. Gordon and R. Swanson at the ICT, University of Southern California, for their assistance in completing this research. This work is funded by the FWF Grant P20269 *TransAgere* and the Know-Center.

4. REFERENCES

- [1] Chirita, P. A., Costache, S., Nejdil, W. and Handschuh, S.: P-TAG: large scale automatic generation of personalized annotation tags for the web, in ‘WWW ’07’, ACM, New York, NY, USA, pp. 845–854, 2007.
- [2] Chulef, A. S., Read, S. J. and Walsh, D. A.: A Hierarchical Taxonomy of Human Goals. In *Motivation and Emotion*, 25 (3), 191–232, 2001.