Reading Between the Lines: Attitudinal expressions in text

Jussi Karlgren, Gunnar Eriksson,
Kristofer Franzén, Preben Hansen
{jussi,guer,franzen,preben}@sics.se
Swedish Institute of Computer Science
Box 1263, 164 29 Kista
Sweden

Stefano Mizzaro
mizzaro@dimi.uniud.it
University of Udine
Via delle Scienze, 20, 33100 Udine
Italy

Paul Clough, Mark Sanderson
{p.d.clough, m.sanderson}@sheffield.ac.uk
University of Sheffield
Regent Court, 211 Portobello St
Sheffield S1 4DP, U.K.

Abstract

This paper describes the starting points of a proposed project for attitude extraction. The project plans to extend language technology methods to text linguistic study, and apply and evaluate results in task-oriented information access systems.

State of the art: words are on lines and we can count them

Information access systems of today share two common assumptions. Firstly, the assumption the information that a reader reads or fetches from a document is sufficiently well represented by the occurrences of terms that make up the text. Secondly, the assumption that the differences in term frequency between documents is sufficient to make reliable judgments about the relative relevance of the documents to a specific information need. Arguably, these starting points are effective. For the typical tasks that information retrieval systems are engaged in, the topicality or the "aboutness" of a document, can be modeled up to a point by terms and their frequencies. Since the 1950's, these assumptions have been paramount in information retrieval system design, with algorithmic variations in implementation (Luhn, 1957); the text analysis is limited to ingesting texts, crunching their terms into tables and discarding as noise extraneous information such as clausal organisation, text style, expressed opinions or sentiments and other less explicitly topical information. This drastic filtering step is arguably a fair starting point for the enterprise of text understanding, but we need a wider perspective on text, topic, and information to be able to even start discussing ephemeral characteristics of text such as quality.

Conversely, in the field of artificial intelligence, text understanding research has focused on top-down processing, on understanding the context of a story, on using prediction of event and causality chains or patterns of previous experience to fill out missing details in narrations. These approaches have not attempted large scale experiments and have not attempted to proceed into the general field of text understanding. The practical utility of the artificial intelligence approaches remains yet to be proven.

While information systems have focused on efficiency and effectiveness, other simultaneous technological advances have lowered the publication threshold dramatically. The attendant explosion in available information is a mixed blessing for consumers — the overall quality of the information available has not necessarily improved at the same pace that the quantity has. The next generation of information access tools must help readers not only find, but also assess and evaluate the pertinence of information made available to them and refine that information to fit the context of the reader.

Beyond Topic: Attitudes and Opinions

Author perspective is indirectly expressed in text by his or her attitudes towards the intended reader, towards the discourse in which the text participates, towards the scene or stage that is delineated in the text together with the players and objects that are involved in it. Language provides mechanisms for expressing attitude, and as human readers we use these mechanisms; information access systems only use a small fraction of them. Authors may express their attitude towards a player or object by attributing a quality by using a certain adjective or express the attitude towards the alleged evidentiality of a certain situation by the choice of a certain verb like *claim* or *deny*. The attitude towards the discourse, the reader, and the text may be expressed by other means, e.g. certain adverbials like *surprisingly*, *arguably* or even *stupidly* or *wonderfully*. But, still, in order to exploit these explicit cues to author attitudes, we need to access the structure of relations between the textual entities mentioned above and the actors that engage in the reading situation, i.e. the author, the reader, the players, the setting. These relations are not always explicitly expressed on lines: to find them we need to read *between the lines*.

Find the players of the narration

We plan to model text not by terms, nor by concepts but by *players* or discourse referents. Discourse referents – a theoretical concept since Coling 1969 (Karttunen, 1969), but hitherto not directly applied to information access technology – introduce a representation of text on a higher level of abstraction than terms are able to, and are text-internally and syntactically detectable – independent of textexternal domain-specific knowledge bases. Identifying potential players in text (as opposed to entities that are mentioned without being players) will need syntactic analysis, at least some initial steps towards anaphora resolution, a theory of topicality in text, and some statistical finesse. We do not aim to push the envelope as regards identification of discourse referents in themselves — the literature on how to identify and formalise discourse referents is plentiful albeit unproven in large scale processing experiments such as the ones we envision (e.g. Grosz et al, 1995; Sidner, 1979 and 1986; Rich and LuperFoy 1988; Fraurud, 1988).

Today we are equipped with better processing tools than previous years (cf. e.g., Tapanainen and Jrvinen, 1997) and we will in this activity use linguistically analyzed textual material to extract examples of what we are interested in. The main focus will be on the identification and classification of lexical noun phrases, i.e., only phrases headed by content-bearing words such as nouns or adjectives will be considered possible manifestations of players at this stage. A combination of syntactic and lexical tools will be employed in the identification task, and for the classification task, statistical methods based on e.g. recurrence and form of the candidate phrase will be employed to select the most likely central referents in the text (cf. Justeson and Katz, 1995); in continuing steps coreference resolution algorithms (cf. e.g. Fraurud, 1988; Lappin and Leass, 1994) will be able to establish that two different lexical noun phrases like “The Swedish prime minister” and “the minister”, and a pronoun ***he*** are referring to the same individual.

As a first step we can exemplify by picking out nouns that enter into a genitive attributive relation to other nouns, such as ***Clinton*** in “Clinton’s recent policy”. Examples from one 1994 month of the Los Angeles Times are given in table 1. Many of them are prime candidates for expression of attitudes: notably the then U.S. President Clinton and the then California Governor Wilson appear on the list (always allowing for some other Clintons and Wilsons to generate some noise in the model). Both can be expected to engender some expression of author attitudes.

153	city	39	school
94	nation	39	Prussia
94	county	32	team
86	California	28	district
82	world	27	Japan
80	state	27	department
68	company	26	region
61	Clifton	25	group
57	woman	23	government
53	country	23	area
50	year	22	man
44	America	20	Wilson
43	administration	20	president
42	today	17	child

Table 1: Nouns that are genitive attributes to other noun phrases in one month of 1994 Los Angeles Times

early	leaving
encouraging	longtime
former	now-famous
standard	opportunistic
actual	outraged
agitated	proposed
entire	real
frequent	regular
gregarious	staunch
high-ranking	underfunded

Table 2: Adjectival attributes to ***Clinton*** in one month of 1994 Los Angeles Times

Find attitudes towards the players

A key to using players – a more abstract level of topical representation than terms – effectively as points of departure for text understanding is to chart the attitudes the text author holds about its players. While players are established in a fairly situation-independent manner, the way players are described and moved on and off stage indicate the tenor, the thrust, and the ecology of a text. This type of analysis has been performed manually in the past for small numbers of texts for the purposes of psychological profiling, political analysis, or unfolding rhetorical structure; later studies in stylistic analysis or authorship identification have methodological parallels.

Our proposed addition to the field of information access is the introduction of a robust and low-key pragmatic component. The sort of questions we will ask of a text include whether discourse entities are mentioned in passing, aggressively, pointedly, irritatedly, with surprise and so forth, and what importance the text as an artifact accords a particular referent? This will require some fine-grained text-syntactic analysis; it will also take a fair amount of generalisation over attributes: the analysis will need to build lexical categories of typical expressions of attitudes as well as touch upon the problems of attribute scoping.

As an example we show in table 2 the adjectival attributes to the noun ***Clinton*** in press text from the Los Angeles Times in 1994. There are some clearly attitudinal adjectives in the lot, better than a set of randomly picked adjective from the same corpus.

Similarly, comparing the set of noun phrases with ***Clinton’s*** as a genitive attribute in table 3 to noun phrases with any genitive attribute in table 4 it is clear that the attitudinal loading of the genitive attribute makes a difference.

Attitudes towards the narration itself

In many texts the expression of attitudes is not clearly directed towards the prominent players of the text, but towards some other entity, e.g. the intended reader or the text itself. Such attitudes are often cues to the author’s intended positioning of the text in a certain discourse situation. Examples are the choice of the already mentioned verbs and adverbs like *deny*, *claim*, *surprisingly* and *arguably*, or more complex phrases and expressions like *the reader might disagree with the position expressed so far*. A number of studies have

Clinton’s white house
Clinton’s strong commitment
Clinton’s proposed alliance
Clinton’s tough talk
Clinton’s proposed reform
Clinton’s prominent role
Clinton’s political quagmire
Clinton’s federal budget
Clinton’s vehement response
Clinton’s strong defense

Table 3: Most frequent heads with genitive attribute ***Clinton*** in one month of 1994 Los Angeles Times

X’s executive director	X’s valuable player
X’s general fund	X’s close friend
X’s good friend	X’s advisory council
X’s general manager	X’s winless streak
X’s central bank	X’s athletic director
X’s young brother	X’s super bowl
X’s general plan	X’s short story
X’s technical program	X’s Greek row
X’s national championship	X’s good player

Table 4: Attribute-head combinations after any genitive attribute in one month of 1994 Los Angeles Times

shown that it is feasible to extract constructions or words like these with machine-learning or corpus-based methods and to use these cues to categorise texts along the negative-positive or subjective-objective dimensions. C.f. (Kushal et al. 2003) — on product reviews, (Pang et al. 2002) — on movie reviews, or (Wiebe, 1994; Wiebe et al. 2001) — on differentiating between objective and subjective passages of text. Based on previous work in this vein and the dynamic construction of lexical databases mentioned above will make it possible to extract some of the attitudes towards the narration or towards referents that are less easy to chart than the ones referred to in the previous section.

Specifically, texts abound with self reference, clause reference, situational references and other types of meta-level references. Examples of such references are the pronoun ***It*** in: *I kissed the ticket collector on the train yesterday.* ***It*** *was nice.* and the pronoun ***That*** in: “*Sometimes there is no correlate.* ***That*** *is an annoying problem.*” Most practically oriented studies on referential expressions gather such cases under the heading “situation reference”. To find out what the pronoun in the example above is referring to is at present problematic or near-impossible, but collecting attitudes expressed towards them is not. In the examples above, we know that the the author regards something as *nice* and something as *annoying*, even if we are unable to identify that entity.

As examples, table 5 gives the most frequent complements to the word ***be*** after ***it***, ***that***, or ***this*** has appeared as a subject; table 6 gives a set of attributes that have comparatively most often appeared with such complements in the same corpus of 1994 Los Angeles Times.

hard	thing	way	good
important	easy	one	be
kind	time	difficult	part
true	something	place	game
possible	case	matter	all
problem	fun	bad	impossible
issue	great	clear	nice

Table 5: Most frequent predicative complements to ***it***, ***that***, or ***this*** in the 1994 Los Angeles Times

once-in-a-lifetime	win-win	unclear
just so	done	sad
just	scary	unfortunate
wishful	judgment	tricky
fitting	wonderful	wake-up
fun	worthwhile	unbelievable

Table 6: Adjectival attributes to predicative complements to ***it***, ***that***, or ***this*** in the 1994 Los Angeles Times

Evaluation and Relevance

Our hypotheses are evaluable by empirical study. The evaluation will be based on analysis of the performance on large collections of text. This can to some extent be done automatically using existing, possibly reannotated corpora, but the most important part will be human assessors judging the efficiency and appropriateness of our methods. In each case, evaluation hinges on the elusive notion of *relevance*.

The concept of relevance lies at the convergence of understanding users, information needs, items of information, and interaction. It ties together all proposed and current research projects in context sensitive information access. Relevance is a function of task, collection characteristics, user preferences and background, situation, tool, temporal constraints, and untold other factors.

In information retrieval research the target concept of relevance is based on the everyday notion, but operationalised to be a relation between query and document. Much of the success of information retrieval as a research field is owed to this formalisation, but today, the strict, abstract, and formalisable relevance of the past decades is becoming something of a bottleneck – since it disregards most non-topical factors it cannot contribute to the evolution of contextually sensitive information access systems.

Relevance can be extended to formally cover non-topical information such as expressed attitudes. It has been proposed to use features of documents (mainly metadata) to exploit beyond-topical facets (Mizzaro 1998). What we will attempt is to relate aspects of relevance to features, especially non-topical ones, extracted from the text.

We will capitalize on previous research on relevance (see e.g. (Mizzaro, 1997; Mizzaro, 1998; Schamber, 1994)) which has emphasized how several different kinds of relevance do exist, and how the “system relevance” implemented in current information retrieval systems is different from the “user relevance”, i.e., the relevance that the final user is interested in. Results from the research on multidimensional relevance and on relevance criteria (Schamber et al., 1990; Barry and Schamber, 1998) has shown during the last decades how beyond-topical factors are used by users to establish the user relevance of a document to an information need.

Harvesting attitudes in texts

The primary objective of our proposed project is to use manifestations of attitudes in text to enrich the representation of a text to expand the possibilities to assess the relevance of the text to a specific reader or information need.

We will enrich the text description by building charts of referents and the attitudes expressed towards them and by gathering attitudes towards the narration. To be able to abstract away from the actual lexical realisations, the attitudes must be typologised in some palette of basic dimensions. Once this is done, the texts themselves can be typologised by characteristics such as "focused", "intensive", "involved", "detached", "positive", "negative", etc.

Our Hypotheses

We believe referents or *players* are the main bearers of topicality in texts. The evidence of our first studies shows that the role of a player is conveyed from author to reader largely by attributes explicitly attached by syntactic mechanisms to the player's occurrences in text. The attitudes towards the text itself is likewise conveyed through the choice of evaluative expressions, and through meta-level references.

We believe expressions introducing, maintaining and evaluating players in text can be identified — even if not fully understood — using the mechanisms published to date. Players are likely to be useful for topically relevant categorisation of texts and will, if evaluated by standard information retrieval evaluation metrics, most likely improve precision at an attendant cost in the nowadays – given large, dynamic data repositories – less crucial measure of recall.

We believe attitudes in texts are heavily dependent on text type and domain, and that experienced readers have learned to understand the systematics of attitudinal mechanisms and employ such categorisations in assessing and reading texts. By categorising texts by attitudinal mechanisms employed in it we believe we are opening the potential for building a helpful tool to close the gap between experienced and less experienced readers. This type of tool will have to be evaluated using new metrics, measuring user confidence in their choice of document and satisfaction with access session rather than the database oriented measures of precision and recall.

References

Barry, C. L. and Schamber, L. 1998. Users' criteria for relevance evaluation: A cross-situational comparison. *Information Processing and Management*, 34:219-236.

Fraurud, K. 1988. Pronoun Resolution in unrestricted text. *Nordic Journal of Linguistics* 11:47-68.

Grosz, B.; Joshi, A; and Weinstein, S. 1995. Centering: a framework for modelling the local coherence of discourse. *Computational Linguistics*, 21:44-50.

Justeson, J.S. and Katz, S.M., 1995. Technical Terminology: Some Linguistic Properties and an Algorithm for Identification. *Natural Language Engineering*, 1:9–27.

Karttunen, L. 1969. Discourse Referents. Reprinted in: McCawley, James D., ed. 1976. *Notes from the Linguistic Underground*, 363-386. New York: Academic Press.

Kushal, D.; Lawrence, S; and Pennock, D.M. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of the Twelfth International World Wide Web Conference. Geneva: IW3C2.

Lappin, S. and Leass, H. J. 1994. An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics* 20:535–561.

Luhn, H.P.. 1957. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development* 1:309-317.

Mizzaro, S. 1997. Relevance: The whole history. *Journal of the American Society for Information Science*, 48:810–832. John Wiley and Sons Inc., New York, NY.

Mizzaro, S. 1998. How many relevances in information retrieval? *Interacting With Computers, Elsevier, The Netherlands*, 10:305–322.

Pang. P.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing . ACL.

Rich, E. and LuperFoy, S. 1988. An architecture for anaphora resolution. In Proceedings of the 2nd Conference on Applied Natural Language Processing, 18-24. Stroudsburg, Pennsylvania: ACL.

Schamber, L. 1994. Relevance and information behavior. In *Annual Review of Information Science and Technology*, 29:3-48.

Schamber, L., Eisenberg, M.B., and Nilan, M.S. 1990. A re-examination of relevance: Toward a dynamic, situational definition. *Information Processing and Management*, 26:755-776.

Sidner, C.L. 1979. Towards a computational theory of definite anaphora comprehension in English discourse. Technical Report No. 537. M.I.T., A.I. Laboratory.

Sidner, C.L. 1986. Focusing in the comprehension of definite anaphora. *Readings in Natural Language Processing*, ed. by B. Grosz, K. Jones and B. Webber. Morgan Kaufmann Publishers.

Wiebe, J. M. 1994. Tracking point of view in narrative. *Computational Linguistics* 20:233-287.

Wiebe, J; Wilson, T.; and Bell, M. 2001. Identifying Collocations for Recognizing Opinions. In Proceedings of ACL 01 Workshop on Collocation. Stroudsburg,Pennsylvania: ACL.