Reading Between the Lines: Attitudinal expressions in text

Jussi Karlgren[1]*   Gunnar Eriksson[1]   Kristofer Franzén[1]   Paul Clough[3]

Preben Hansen[1]   Stefano Mizzarro[2]   Mark Sanderson[3]

October 3, 2003

Abstract

This is a brief overview of the starting points a project currently proposed and under evaluation by funding agencies. We discuss some of the linguistic methodology we plan to employ to identify and analyze attitudinal expressions in text, and touch briefly on how to evaluate our future results.

1   Reading Between the Lines — a project proposal

The proposed project outlined here has at its starting points the extension of current state of the art in language technology to text linguistics and the application area of information access. We are thinking about how to better analyze the information in a textual information repository and how to present the findings of an automatic information access system to its users — in ways that are pertinent to the task at hand.

2   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 (e.g. Karlgren, 2000).

This drastic filtering step is arguably a fair starting point for the enterprise of text understanding. We all know texts are more than bags of words. 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. However, these approaches have not attempted large scale experiments and indeed have not — largely due to lack of efficient linguistic analysis mechanisms — attempted to proceed into the general field of text understanding for some purpose beyond modeling or simulation of cognitive processes. 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

---

*Contact author: jussi@sics.se
[1]SICS - Swedish Institute of Computer Science
[2]Università di Udine
[3]Sheffield University

available to them and refine that information to fit the context of the reader. This is especially crucial, useful, and fruitful in an European context, with numerous languages and a large population of bi- and multi-lingual readers. This proposed project will address one such aspect of the information potential of a document.

3   Beyond Topic: Attitudes and Opinions

One valuable source of information hitherto unused by information retrieval systems is the author's perspective when writing the document. The perspective of the author is indirectly expressed in the 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. To some extent the expression of these attitudes is found on the lines of the text. Text is a one-dimensional structure which attempts to encode the very definitely multi-dimensional cognitive structures the author wishes to transmit to the audience. Text provides linguistic mechanisms for this purpose: the text is a web of entities and relations between those entities encoded as predications and relational expressions of various kinds. As humans we use these mechanisms; information access systems and other text understanding systems of today 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 expressed on lines: to find them we need to read *between the lines*.

4   What can we do about it?

Today many earlier bottlenecks have been overcome. Linguistic analysis technology has developed: tools for linguistic analysis are at hand and the challenge is now not to find methods to extract linguistic information about sentences in texts, but to figure out what of that information to use, for what purposes, how to refine it further, and how to find it reliably.

On the other hand, the methodological end, this project intends to contribute to an improved perspective. It is impossible to fully predict in advance what interpretation a reader is going to arrive at, but – and this is the assumption on which this proposal rests – it is feasible to better approximate the information the reader will perceive by the use of non-topical information such as the attitudes expressed in the text. Furthermore, we believe that this can be achieved by an approach that makes use of new outlooks on combinations of standard methods in text and computational linguistics, and finally, we assume that the obtained results will enable us to discuss and refine the notion of a text's relevance to a reader.

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 text-external 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 Järvinen, 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

Hypothesis 1: Players referents or *players* are the main bearers of topicality in texts and expressions introducing and maintaining them in text can be identified — even if not fully understood — using the mechanisms published to date. Players can be used for topical categorisation of texts if evaluated by standard information retrieval evaluation metrics and will improve precision at an attendant cost in the nowadays – given large, dynamic data repositories – less crucial measure of recall.

Our belief is that we will be able to achieve practically applicable and academically noteworthy results with small effort for several languages and for large data repositories.

5 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. Today, with reliable and robust dependency grammatical analysis available (Tapanainen and Järvinen, 1997) and recent advances in dynamic construction of lexical databases (Karlgren and Sahlgren, 2001) we have a technological base analysis beyond immediate context.

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.

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

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

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

Table 4: Attribute-head combinations after any genitive attribute in one month 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

6   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 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.

Hypothesis 2: Attitudes Human assessors are able to agree on the attitudes of the author vis a vis discourse players in text.

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.

Human assessors are also able to agree on the attitudes of the author towards the text itself.

The attitudes towards the text is conveyed through the choice of evaluative expressions, and through meta-level references.

All the mechanisms above can be identified and analyzed using analysis tools available today.

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.

Hypothesis 3: Text types Attitudes in texts are heavily dependent on text type and domain. Thus, texts can be categorised according to the attitudinal mechanisms employed in it.

Proficient readers have learned to understand the systematics of attitudinal mechanisms and employ such categorisations in assessing and reading texts.

7   Evaluation and Relevance

The hypotheses we have postulated are couched in evaluable terms. 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. We will discuss and establish how relevance can be extended to formally cover non-topical information such as expressed attitudes and formulate experimental goals to verify the new formalisations.

Very little has been done in this direction so far. In the past, it has been proposed to use features of documents (mainly metadata) to exploit beyond-topical facets (Mizzaro 1998). What is novel in this proposal, and has never been attempted before, is to aim at beyond-topical relevance by exploiting language technology techniques to extract features of the text.

We will capitalize on previous research on relevance. This large body of research (see [Mizzaro, 1997, Schamber, 1994] for the most recent reviews) has emphasized how several different kinds of relevance do exist [Mizzaro, 1998], 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.

Hypothesis 4: Relevance Categories of text are to a differing degree useful for various tasks and various populations of reader. This can be expressed in terms of an extended notion of relevance.

8   This is not trivial, but may be useful

With an exploitation of discourse referents and a move towards a more semantic-based and contextually aware text structure, the research ideas of our proposed project are a new approach in a long series of unsuccessful attempts at applying language technology to information retrieval. Research has shown that despite decades of advances in language technology, its application to the retrieval process, for retrieval over general domain texts, has provided little or no benefit over and above what can be achieved using statistically-based word counting methods. Applications of NLP techniques, such as indexing documents by phrases (Fagan, 1987), grammatical word tagging (Sacks-Davis, 1990; Smeaton, 1992), word sense disambiguation (Sanderson, 1994), and sentence parsing (Smeaton and Sheridan, 1991) have often proved to harm retrieval effectiveness in vanilla information retrieval tasks. Even if the technology development from such research is successful, when deployed, the new services tend to be ignored by searchers (Anick, 2003) as they are perceived to be of little practical value.

The situation is such that for many of the top performing systems in the TREC annual benchmarking of retrieval algorithms (http://trec.nist.gov/), the only application of relatively sophisticated linguistic knowledge to be found is the use of stemmers, which normalise word forms, (Porter, 1980; Krovetz, 1993). A similar dearth of language technology is believed to prevail in major search engines such as Google.

Based on past performances of language technology and information retrieval, one can conclude that our proposed project will be a high risk venture. Unlike many of the past approaches, however, we believe that the risk will be confined to proving the technology of attitude identification; if successful, we contend

that the take up of the technology by technology developers and users will be great. We envision its use in tasks beyond the most simple information retrieval use cases — tasks where more depth of information is necessary to perform well.

Hypothesis 5: Use Attitudinal information and mechanisms to detect and identify such information can be applied in the development of practical tools.

Attitudinal information is one piece of a puzzle to add finer grain to the simple picture of text and information that today's information access systems operate with. It is an efficient picture but with too little relief for readers to make the multi-faceted and oftentimes complex decisions expected of them in an information society.

References

[Anick, 2003] Anick, P. (2003) Using terminological feedback for web search refinement: a log-based study. Proceedings of the 26th annual international ACM SIGIR conference on Research and Development in Information Retrieval: 88-95

[Barry and Schamber, 1998] Barry, C. L. and Schamber, L. (1998). Users' criteria for relevance evaluation: A cross-situational comparison. *Information Processing and Management*, 34(2/3):219–236.

[Fagan, 1987] Fagan, J. (1987): Experiments in automatic phrase indexing for document retrieval: a comparison of syntactic and non-syntactic methods, in Doctoral Dissertation, Technical Report TR 87-868

[Fraurud, 1988] Fraurud, Kari. 1988. "Pronoun Resolution in unrestricted text". *Nordic Journal of Linguistics* 11, 47-68.

[Grosz et al, 1995] Grosz, B., A. Joshi and S. Weinstein. 1995. "Centering: a framework for modelling the local coherence of discourse". *Computational Linguistics*, 21(2), 44-50.

[Justeson and Katz, 1995] Justeson, J.S. and Katz, S.M., 1995. "Technical Terminology: Some Linuistic Properties and an Algorithm for Identification in Text". *Natural Language Engineering*, 1:9–27.

[Karlgren, 2000] Karlgren, Jussi. 2000. *Stylistic Experiments in Information Retrieval*. Ph D Dissertation. Stockholm: Stockholm University.

[Karlgren and Sahlgren, 2001] Karlgren, Jussi and Sahlgren, Magnus. 2001. "From Words to Understanding". In Uesaka, Y., Kanerva, P. and Asoh, H. (Eds.): *Foundations of Real-World Intelligence*, pp. 294-308, Stanford: CSLI Publications.

[Karttunen, 1969] Karttunen, Lauri. 1969. "Discourse Referents". Reprinted in: McCawley, James D., ed. 1976. *Syntax and Semantics 7: Notes from the Linguistic Underground*, pp. 363-386. New York: Academic Press.

[Krovetz, 1993] Krovetz, R. (1993): Viewing morphology as an inference process, in Proceedings of the 16th annual international ACM SIGIR conference on Research and Development in Information Retrieval: 191-202

[Kushal et al, 2003] Kushal, Dave, Steve Lawrence and David M. Pennock. "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews". *Proceedings of the Twelfth International World Wide Web Conference (WWW-2003)*, May 2003.

[Lappin and Leass, 1994] Lappin, Shalom and Herbert J. Leass. 1994. "An Algorithm for Pronominal Anaphora Resolution." *Computational Linguistics* 20 4:535–561.

[Luhn, 1957] Luhn, Hans Peter. 1957. "A Statistical Approach to Mechanized Encoding and Searching of Literary Information." *IBM Journal of Research and Development* 1 (4) 309-317.

[Mizzaro, 1997] Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science*, 48(9):810–832. John Wiley and Sons Inc., New York, NY. Republished in "Historical Studies in Information Science", T. Bellardo Hahn e M. Buckland editors, 1998, ISBN:1-57387-062-5.

[Mizzaro, 1998] Mizzaro, S. (1998). How many relevances in information retrieval? *Interacting With Computers, Elsevier, The Netherlands*, 10(3):305–322. ISSN: 0953-5438. Paper awarded with the Informer 'Best Student Paper in IR' award.

[Pang et al, 2002] Pang. Bo, Lillian Lee, Shivakumar Vaithyanathan. 2002. "Thumbs up? Sentiment Classification using Machine Learning Techniques". *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[Porter, 1980] Porter, M.F. (1980): An algorithm for suffix stripping, in Program - automated library and information systems, 14(3): 130-137

[Rich and LuperFoy, 1988] Rich, Elaine and Susann LuperFoy. 1988. "An architecture for anaphora resolution". *Proceedings of the Second Conference on Applied Natural Language Processing (ANLP-2)*, 18-24, Texas, U.S.A.

[Sacks, 1990] Sacks-Davis, R., Wallis, P., Wilkinson, R. (1990): Using syntactic analysis in a document retrieval system that uses signature files, in Proceedings of ACM SIGIR Conference, 13: 179-191

[Sanderson, 1994] Sanderson, M. (1994) "Word sense disambiguation and information retrieval", Proceedings of the 17th ACM SIGIR Conference, Pages 142-151

[Schamber, 1994] Schamber, L. (1994). Relevance and information behavior. In *Annual Review of Information Science and Technology*, volume 29, pages 3–48.

[Schamber et al., 1990] 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(6):755–776.

[Sidner, 1979] Sidner, Candy L. 1979. "Towards a computational theory of definite anaphora comprehension in English discourse". Technical Report No. 537. M.I.T., Artificial Intelligence Laboratory.

[Sidner, 1986] Sidner, Candy 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.

[Smeaton and Sheridan, 1991] Smeaton, A.F., Sheridan, P. (1991) "Using morpho-syntactic language analysis in phrase matching" Proceedings of RIAO 91, Intelligent Text and Image Handling: 414-429

[Smeaton, 1992] Smeaton, A. (1992) "An evaluation of retrieval performance using simple statistics and SIMPR linguistic processing on a standard collection of texts" SIMPR technical report (SIMPR-DCU-1992-50.2i)

[Tapanainen and Järvinen, 1997] Tapanainen, Pasi and Timo Järvinen, 1997. "A non-projective dependency parser". *Proceedings of the 5th Conference on Applied Natural Language Processing*, pp. 64-71. March 31st - April 3rd, Washington D.C., USA.

[Wiebe, 1994] Wiebe, Janyce M. 1994. "Tracking point of view in narrative". *Computational Linguistics* 20 (2): 233-287.

[Wiebe et al, 2001] Wiebe, Janyce, Wilson, Theresa, and Bell, Matthew. 2001. "Identifying Collocations for Recognizing Opinions". *Proc. ACL 01 Workshop on Collocation*. Toulouse, France, July 2001.