

Improving Relevancy Accessing Linked Opinion Data

Boris Galitsky¹, Josep Lluís de la Rosa² & Gábor Dobrocsi³

¹ Univ. Girona Spain

bgalitsky@hotmail.com

² EASY Innovation Center, Univ Girona Spain

pepluis@eia.udg.edu

³ Univ Miskolc Miskolc Hungary

gadomail@gmail.com

Abstract

We introduce a search engine and information retrieval system for providing access to linked opinion data. Natural language technology of generalization of syntactic parse trees is introduced as a similarity measure between subjects of textual opinions to link them on the fly. Information extraction algorithm for automatic summarization of web pages in the format of Google sponsored links is presented. We outline the usability of the implemented system, integrated opinion delivery environment (IODE), box.cs.rpi.edu:8080/wise/lancelot.jsp

Introduction

Nowadays, it is becoming more and more popular to enable people to share structured data on the Web. Design of web portals leverages the fact that value and usefulness of data increases, when the degree of inter-links with other data rises. It is especially true for opinion data, where trust to an aggregated opinion can be developed by a demonstration of a highly interlinked sources of data of various modalities (Kim and Hovy 2004, Aciar et al 2007). The sources of opinion data include reviews on products and services, submitted to respective retailer and provider websites, opinion sharing portals, blogs and forums.

In this paper we introduce integrated opinion delivery environment (IODE), which provides a user with opinions and recommendations about products and services expressed in a wide spectrum of forms, from digested, aggregated and abbreviated authentic user opinions to advertisement form. Furthermore, we will link the conventional opinion data with implicitly expressed product opinions extracted from wireless location services. For a particular business location, we show the area most loyal customers to this business reside.

IODE includes search for products and services by names, parameters and user needs. The result is displayed as integrated opinions from various sources, as well as

automatically generated advertisement from original web page of product provider, following the style of Google sponsored links, which is natural for search engine users.

Obviously, automatically linking textual and other forms of data on the fly requires some kind of simplified text matching techniques, so that the portions of textual data which is linked is *relevant to each other* and *to the user query*. To link textual data, one needs to come up with a robust measure of semantic similarity between text fragments. In this paper we develop a domain-independent measure of text relevancy based on machine learning of syntactic parse tree.

To make opinion sharing easier and more efficient, we aim at coherent structure of opinion expression for both positive and negative sentiments. We select Google sponsored link format as a basis for opinion sharing. To encourage both business owner / advertiser and user to express their opinion in this form, we need a hybrid of information extraction and summarization techniques to extract expressions suitable to form advertisement line from a business web page.

System description

The system is designed to provide opinions data in linked aggregated form obtained from various sources. Conventional search results and Google sponsored link formats are selected as most effective and already accepted by the vast community of users.

User interface of IODE is shown at Fig.1. To search for an opinion, a user specifies a product class, a name of particular products, a set of its features, specific concerns, needs or interests. A search can be narrowed down to a particular source; otherwise multiple sources of opinions (review portals, vendor-owned reviews, forums and blogs available for indexing) are combined.



 site(optional): lang: en ▾

You are probably better off buying a new camera. The only... - Cameras ...
 ... market or is it better to just buy a good **digital camera** and a **waterproof case**? ...
 If this seems excessive, then a **plastic case** that goes around a point & shoot ...
<http://futureshopforums.com/futureshop/board/message?message.uid=116119>

Single question and answers : Yahoo! Tech
 ... Fujifilm, which is essentially a disposable **camera** in a sealed **plastic case**. ... Finally,
 the most expensive option: a new **digital camera** that's **waterproof**. ...
<http://tech.yahoo.com/qa/20090320071656AAGV8tm>

Kid Tough Digital Camera
 [Kid-Tough **Digital Camera Case** - Blue] 79 results, prices starting at \$1, Compare and
 Save. ... Kid-Tough Pink **Waterproof Digital Camera** ...
<http://www.shopwiki.com/Kid+Tough+Digital+Camera>

Fig. 1: User interface of Integrated Opinion Delivery Environment.

Opinion search results are shown on the bottom-left. For each result, a snapshot is generated indicating a product, its features which are attempted by the system to match user opinion request, and sentiments. In case of a query which includes multiple sentences, a hit contains combined snapshot of multiple opinions from multiple sources, dynamically linked to match user request.

Automatically generated product advertisements compliant with Google sponsored links format are shown on the right. Phrases in generated advertisements are extracted from original product web pages and possibly modified for compatibility, compactness and appeal to potential users. There is a one-to-one correspondence between products in opinion hits on the left and generated advertisements on the right (unlike in Google, where sponsored links list different websites).

When a hit is selected by a user, for a local business, we show the areas of residence of users for this business. The data is obtained from wireless location based services, recording signal strength for each cell phone at each moment Galitsky and Miller 2006). Given the signal strengths from a number of cell towers at a moment, one can determine user locations such as a shopping mall, particular shop, or a car dealership. For each business, we build offline an index of associations between locations of businesses and residences of the users of this business. In search time, we show at Google map the areas of residence of the customers who have physically come to this business location.

Both respective business representatives and product users are encouraged to edit and add advertisements, expressing product feature highlights and usability opinions respectively. This features assures openness, objectivity, and authenticity assured by community



Sponsored links			
Good underwater digital camera Go 10 - 20 feet down in saltwater	Good digital camera and Think up are Pentax Optio 10MP Digital Camera	Slr camera but Say is that Olympus in Japan was very nice about servicing it	New camera Be very careful to follow all instructions provided by Olympus do the checks
Canon digital camera Please enable your browser My Tech column	My Canon point and Check out httpwww Look at the Olympus 1030SW	Also buy generic underwater soft Try the Pentax W60 Other Yahoo	My Tech Find out more at Water park this year and
Fisher Price Kid Tough Memorex Dora Kid Tough Digital Camera Model 01124			

participation in providing access to linked opinions for other users. For example, a negative experience staying in a hotel can be expressed like

*Disney World Super 12 Motel
 Safest place in the area*

Take your car there and have it stolen

Search phrase may combine multiple sentences, for example: *"I am a beginner user of digital camera. I want to take pictures of my kids and pets. Sometimes I take it outdoors, so it should be waterproof to resist rain"*.

Obviously, this kind of specific opinion request can hardly be represented by keywords like 'beginner digital camera kids pets waterproof rain'. So a semantic search engine rather than a keyword-based one is required for such task. For a multi-sentence query the results are provides as linked search hits:

Take Pictures of Your Kids? ... Canon 400D EOS Rebel XTi **digital SLR camera** review ↔ I am by no means a professional or long time user of SLR cameras.

How To **Take Pictures Of Pets And Kids** ... Need help with **Digital slr camera** please!!!! - Yahoo! Answers ↔ I am a **beginner** in the world of the **digital SLR** ...

Canon 400D EOS Rebel XTi **digital SLR camera** review (Website Design Tips) / Animal, **pet, children**, equine, livestock, farm portrait and stock ↔ I am a **beginner** to the slr **camera** world. ↔ I want to **take** the best **picture** possible because I know you. Call anytime.

Linking (↔) is determined in real time to address each part in a multi-sentence query which can be a blog posting seeking advice. Linked search results are providing comprehensive opinion on the topic of user interest, obtained from various sources and linked on the fly.

Generalizing a pair of sentences

To measure of similarity of abstract entities (Cardie & Mooney 1999) expressed by logic formulas, a least-general generalization was proposed for a number of machine learning approaches, including explanation based learning and inductive logic programming. In this study, to measure similarity between natural language (NL) expressions, we extend the notion of generalization from logic formulas to syntactic parse trees of these expressions. If it were possible to define similarity between natural language expressions at pure semantic level, least general generalization would be sufficient. However, in horizontal search domains where construction of full ontologies for complete translation from NL to logic language is not plausible, therefore extension of the abstract operation of generalization to syntactic level is required. Rather than extracting common keywords, generalization operation produces a syntactic expression that can be semantically interpreted as a common meaning shared by two sentences (compare with Bar-Haim et al 2005, Hacioglu 2004).

The purpose of an abstract generalization is to find commonality. Generalization operation occurs on the following levels:

- Text
- Paragraph
- Sentence
- Phrases (noun, verb and others)
- Individual Word

At each level except the lowest one of Individual Words, the result of generalization of two expressions is a *set* of expressions. In each such set, the expressions for which there exist less general expressions are eliminated. Generalization of two sets of expressions is a set of sets which are the results of pair-wise generalization.

We first outline the algorithm for two sentences and then proceed to the specifics for particular levels.

- 1) Obtain parsing tree for each sentence. For each word (tree node) we have lemma, part of speech and form of word information, as well as an arc to the other node.
- 2) Split sentences into sub-trees which are phrases for each type: verb, noun, prepositional and others; these sub-trees are overlapping. The sub-trees are coded so that information about occurrence in the full tree is retained.
- 3) All sub-trees are grouped by phrase types.
- 4) Extending the list of phrases by adding equivalence transformations (Section 3.2).
- 5) Generalize each pair of sub-trees for both sentences for each phrase type.
- 6) For each pair of sub-trees yield the alignment, and then generalize each node for this alignment. For the obtained set of trees (generalization results), calculate the score.

- 7) For each pair of sub-trees for phrases, select the set of generalizations with highest score (least general).
- 8) Form the sets of generalizations for each phrase types whose elements are sets of generalizations for this type.
- 9) Filtering the list of generalization results: for the list of generalization for each phrase type, exclude more general elements from lists of generalization for given pair of phrases.

For a given pair of words, only a single generalization exists: if words are the same in the same form, the result is a node with this word in this form. We refer to generalization of words occurring in syntactic tree as *word node*. If word forms are different (e.g. one is single and other is plural), then only the lemma of word stays. If the words are different but only parts of speech are the same, the resultant node contains part of speech information only and no lemma. If parts of speech are different, generalization node is empty.

For a pair of phrases, generalization includes all *maximum* ordered sets of generalization nodes for words in phrases so that the order of words is retained. In the following example

To buy digital camera today, on Monday

Digital camera was a good buy today, first Monday of the month

Generalization contains { *digital - camera , today - Monday* } , where part of speech information is not shown. *buy* is excluded from both generalizations because it occurs in a different order in the above phrases. *Buy - digital - camera* is not a generalization because *buy* occurs in different sequence with the other generalization nodes.

As one can see, multiple maximum generalizations occur depending how correspondence between words is established, multiple generalizations are possible. In general, totality of generalizations forms a lattice. To obey the condition of maximum we introduce a score on generalization. Scoring weights of generalizations are decreasing, roughly, in following order: nouns and verbs, other parts of speech, and nodes with no lemma but part of speech only. In its style generalization operation follows along the lines of the notion of 'least general generalization', or anti-unification if a node is a formula in a language of logic. Hence we can refer to the syntactic tree generalization as the operation of *anti-unification of syntactic trees*.

Result of generalization can be further generalized with other parse trees or generalization. For a set of sentences, totality of generalizations forms a lattice: order on generalizations is set by the subsumption relation and generalization score. Generalization of parse trees obeys the associativity by means of computation: it has to be verified and resultant list extended each time new sentence is added. Notice that such associativity is not implied by our definition of generalization.

Generating advertisement from a webpage

IODE builds 3-line advertisements (ads) in a specific format to mimic ads for search engine paid advertisement. Online adverts always link to a specific page. This page may be a homepage to a site offering products or services, or a specific page within that site or a page constructed specifically to be linked to from the ad. Features of products or services as described on the landing page are extracted along with 'advertising' language such as positive sentiment and calls to action. The formulation of ad relies on appropriate product/service descriptions from the landing page. Using a database of existing adverts on Google, the system is capable of finding the closest ad and adjusting it to the web page.

Information extraction problem is formulated as extracting and modifying three expressions from a web pages such that these expressions:

- serve as a page abstract, are as close as possible to the content of the page. Every line of an ad contains one of the important messages from the webpage, and may include the name of brands, companies and products from this page.
- obey the conventional ad style: certain meaning is shared between them, one of these expression should be imperative, all of them should be positive and mention important well understood parameters.

To achieve the above criteria, we combine two following technique implemented as parallel components:

- Syntactic information extraction (SIE); it extracts portions of text and modifies them to adhere to a typical advert phrasing style. This component assures that resultant advert lines are sufficiently closed to original webpage content.
- Template-based (TB); it finds a series of existing adverts mined on the web, which are semantically close to the given webpage, and modifies them to form original advert with the lines matching the content of this webpage. This component assures that resultant advert is a typical advert in terms of phrasing.

Combination of these components assures that the content is well represented and also well phrased in the resultant advert.

For example, from the content like

At Barclays **we believe in great loan deals, that's why we offer 9.9% APR typical on our loans of £7,500 to £25,000****. It's also why we pledge to pay the difference if you're offered a better deal elsewhere.

What you get with a personal loan from Barclays:

* An instant decision if you're an Online Banking customer and **get your money in 3 hours**, if accepted†

* Our price guarantee: if you're offered a better deal elsewhere we'll pledge to pay you the difference between loan repayments***

* **Apply to borrow up to £25,000**

* No fees for arrangement or set up

* Fixed monthly payments, so you know where you are
* Optional tailored Payment Protection Insurance.

We want to generate ads

Great Loan Deals

9.9% APR typical on loans of
£7,500 to £25,000. Apply now!

Apply for a Barclays loan

We offer 9.9% APR typical

Get your money in 3 hours!

Syntactic information extraction

Syntactic information extraction is a rule-based deterministic system implemented as a state machine. Rules include semantic template consisting of individual words and word forms including part of speech constraints. There are rules for beginning of extraction, termination of extraction, and acceptance of extraction. These semantic templates serve as evidence that an expression. For example, semantic template in the beginning of extraction rules verifies the evidence if expression of interest is about to follow. For example: "In our <business_name [noun phrase]> you should" is an instance of a beginning of extraction rule (verb phrase of recommended activity is to follow), "and you" is an instance of termination of extraction rule (next verb group is to follow, so the current one has ended), and "-not"<is excluded> is an instance of acceptance rule (meaning of a negation can be ambiguous, so it is safer to reject current expression). One more example of a beginning rule: if the previous word is 'we' and the current word is 'sell', convert the expression into *get <something>* from the original expression '*we sell <something>*', where *<something>* is unchangeable expression to be included in the resultant advert line.

The procedure of semantic extraction implements the state machine (Fig.2, Galitsky 2003), where each state is characterized by a current extraction part and position in text. Deterministic extraction rules take into account these state parameters and perform transition to the next state (iterating to the next word), changing verification of parameters for this word accordingly. For example, if noun phrase extraction is being selected, we search for the part of speech which is neither noun, adjective, adverb or number, checking occurrence of a web page entity. If an imperative expression is being extracted, we search for the end of verb phrase. For the sake of avoidance of the conflict between different types of extractions and system performance, various types of extractions occur in parallel. Main types of expressions are imperative verb phrases and noun phrases. In the context of ad generation we refer to the class of sentences encouraging potential customers to perform an action, be it physical or mental (epistemic), as *imperative expressions*. The start of imperative expression is either an imperative verb ('sign up for ...'), or an expression indicating that certain activity is expected to be performed by a user ('*we allow immediate cash withdrawal*'). In the latter case we reformulate the expression to derive an explicit imperative: ('get

immediate cash withdrawal', 'take advantage of immediate cash withdrawal'). We use an explicit list of verbs that are relevant to product web pages; Proper position for imperative is either beginning of sentence or a word which follows the one which is determined as an end of a potential previous extraction

There are following constraints for a noun group which includes product entity:

- 1) Such entity should be derived from important webpage components such as title and/or keyword lists
- 2) Such entity should have a modifier or be combined with another noun group which express additional constraint for the main one including entity ('auto-focus digital camera with flat-screen LCD', 'Type3 USB slot for 2GB memory card').
- 3) This group should have more than 2 words, otherwise it is usually too brief to form a headline.
- 4) To have a higher weight, a noun phrase with entity should be under the focus of sentiment, webpage authors are expressing a positive sentiment about the product and/or its particular feature) .
- 5) Extraction of an entity-based noun group is crossed check against a 'conventional' linguistic noun groups (in a narrower sense) obtained by chunking systems.

Building ads from expressions

Above we described how to extract an individual advert line from a webpage, and in this section we proceed to the algorithms of how adverts are formed from individual lines.

Each line belongs to the following class:

1. Noun phrase which includes an entity with modifiers which is important for this page, possibly including product names, brand, or important component of product, such as 'Sony digital camera' and 'LCD'. Positive sentiments can be added to such noun groups: *'inexpensive car insurance'* ('inexpensive' is domain-independent), *'bright overhead projector'* ('bright' here is domain dependent, specific to projectors) .
2. Imperative call to action. This is a wide class of expressions requiring special extraction treatment , based on verbs. This class is wider than just containing imperative verbs; it can be referred to from a wider angle of 'semantic imperative': explicit or implicit call for an action to by a product or to enroll in a service. This expression either starts from an imperative verb, or is converted into imperative form from "we sell..., you can..., we allow".
3. Verb phrase including entity, denoting important actions and advantages of products displayed, like *'use WiFi to explore new ways to connect'*. For verb phrases, special groups of verbs serve different purposes in respect to forming advert lines. Some verbs such as mental actions are not included in

resultant lines, whereas some domain – independent with universal meaning ('use', 'amplify') and domain-specific verbs such as 'amplified'.

Reserved expression. We form classes of equivalents for expressions applicable to wide classes of products such as {'20 in stock', 'currently available', 'available now'} => 'in stock'. We would have to extract delivery information and stock information

Forming an ad, we substitute an ad template with the lines of certain class:

Headline (class1): <positive sentiment><entity> OR just <Entity>
 Line1: <semantic imperatives (class 2) and statement about features (class 3)> (taken from VP chunks using above rules)
 Line2: <call to action, class3 or class4> (either imperative from page OR we reserved expression: "Buy Now! Only <extracted price>" | "In Stock now! Next day shipping" | "Buy Now! Worldwide delivery").

It is important to balance properly between the roles for line of each class, to maximize the number of advert derived given available set of lines.

Conclusions

Developed system outlined in this paper leverages a suite of linguistic and machine learning techniques to link textual opinion data online. In addition to showing opinions as abstracts of search results, the conventional format of search engines, we also show search result content as in Google-like advertisement format. We link the latter to the former to provide more comprehensive and decision-supportive opinion search experience. In this study we do not focus on extracting opinions from text is a subject of multiple studies (Aciar et al 2007, and our paper (Galitsky et al 2009b). Evaluation of text learning based on generalization is also available (Galitsky et al 2009a).

This system can be viewed as an open advertisement network where the advertisements can be edited and extended by both providers and consumers with positive as well as negative sentiment. We expect such open advertisement network to provide more trusted opinion data than a conventional one with paid advertisement model and central control of user impression. We believe that when the corpus of opinion data grows and become more and more important for shopping decision making, distributed and open opinion network will demonstrate a superior performance over the centrally-controlled conventional advertisement network. Proposed approach to link opinion data and modify it by a community can be viewed as a wiki-like approach.

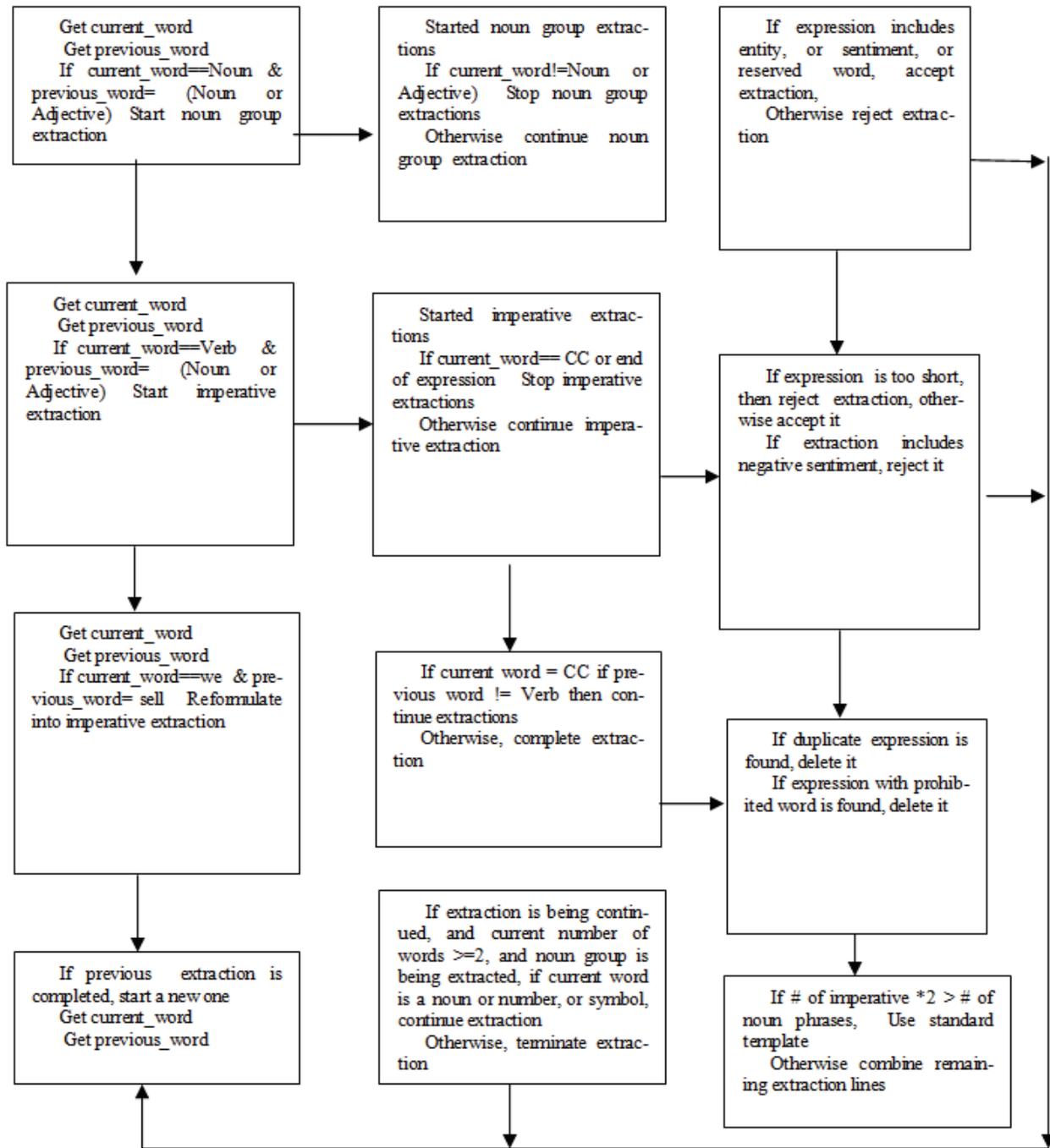


Fig. 2: State machine for syntactic information extraction.

The presented system is designed for end users, however one of its important targets is small and medium size business owner who currently do not use search engine marketing and paid Google advertisement. For this category of users the developed system is expected to be a first introduction to SEM. Searching for their own web pages, they will see auto generated advertisement and access the click data. This first experience will hopefully

take them to the next step of search engine marketing of their products. Hence the system like presented in this paper is focused on tail category of products.

We refer to IODE, box.cs.rpi.edu:8080/wise/lancelot.jsp as *the largest* virtual advertisement network because the search allows to see advertisement generated for all products, not just for those submitted to search engines via a marketing campaign. Most tail products are not represented in Google sponsored links but are represented in IODE, therefore joining the advertisement ‘club’ on a basis of unbiased opinions of business owners and

customers. We observed that search relevance, which is improved by semantic means for conventional data (Thompson et al 1997, Durme et al 2003), is essential for the proposed advertisement model. We specifically tuned the search engine to process complex opinion-related queries. Moreover, search engine employs a special co-reference model to treat multiple sentences in queries. The ad generation system work for arbitrary domains, though some tuning to specific domains may be occasionally appropriate to improve performance. Domains we have tested include retail shopping sites, universities, government, debt management and banks. This is possible because information extraction and summarization technique is based on syntactic rules and machine learning, so domain-specific ontologies are not required.

Recommendations on Consumer Product Reviews," IEEE Intelligent Systems, vol. 22, no. 3, pp. 39-47, May/June 2007.

References

- Bar-Haim, R., Dagan, I., Greental, I. Shnarch, E. Semantic Inference at the Lexical-Syntactic Level AAAI-05.
- Hacioglu, K., Pradhan, S., Ward, W., Martin, J. H. and Jurafsky D. 2004. Semantic role labeling by tagging syntactic chunks. In *Proc. of CoNLL-04*, 2004.
- Kim, S.-M. and Hovy, E. Determining the sentiment of opinions. In *COLING-2004*, pp 1367–1373, Geneva, Switzerland (2004).
- Cardie, C., Mooney R.J. Machine Learning and Natural Language. *Machine Learning* 1(5), 1999.
- Galitsky, B. Natural Language Question Answering System: Technique of Semantic Headers. *Advanced Knowledge International*, Australia 2003.
- Galitsky, B., Kuznetsov SO, Neznanov AA. Inferring semantic properties of sentences mining syntactic parse trees. *ICCS-09 Workshop on Conceptual Structures for Extracting Natural Language Semantics*, Moscow, Russia July 2009a.
- Galitsky, B. Huanjin Chen, Shaobin Du. Inversion of forum content based on authors' sentiments on product usability AAAI SSS-09 Symposium "Social Semantic Web: Where Web 2.0 Meets Web 3.0" 2009b.
- Galitsky, B. and Miller, A. Determining possible criminal behaviour of mobile phone users by means of analysing the location tracking data. *AAAI Spring Symposia on Homeland Security* Stanford CA Tatu, M., and Moldovan, D. 2006. A logic-based semantic approach to recognizing textual entailment. In *Proceedings of the COLING/ACL*.
- Durme, B. V.; Huang, Y.; Kupsc, A.; and Nyberg, E. 2003. Towards light semantic processing for question answering. *HLT Workshop on Text Meaning*.
- Thompson, C., Mooney, R., and Tang, L. 1997. Learning to parse NL database queries into logical form. In *Workshop on Automata Induction, Grammatical Inference and Language Acquisition*.
- Aciair S., Debbie Zhang, Simeon Simoff, John Debenham, "Informed Recommender: Basing