Analyzing Microtext: Papers from the 2011 AAAI Workshop (WS-11-05)

Learning Ontologies from the Web for Microtext Processing

Boris A.Galitsky, Gábor Dobrocsi, and Josep Lluis de la Rosa
University of Girona Spain
bgalitsky@hotmail.com, gadomail@gmail.com, peplluis@isac.cat

Abstract

We build a mechanism to form an ontology of entities which improves a relevance of matching and searching microtext. Ontology construction starts from the seed entities and mines the web for new entities associated with them. To form these new entities, machine learning of syntactic parse trees (syntactic generalization) is applied to form commonalities between various search results for existing entities on the web. Ontology and syntactic generalization are applied to relevance improvement in search and text similarity assessment in commercial setting; evaluation results show substantial contribution of both sources to microtext processing.

Introduction

In recent years, processing short fragment of unstructured text (microtext) became an important area in social content, content aggregation, search and recommendations. Microtext became a popular communication media, and a lot of valuable data is not available in full text format, for which a wide range processing means are currently available. The main bottleneck for processing short fragments of text is a lack of statistical data, therefore deeper linguistic processing and ontology-based methods are expected to substitute statistical ones. One cannot estimate a significance of a keyword in a blog posting by its frequency, so without having an ontology entry for this keyword, it is hard to adequately process it.

It is well known that building, tuning and managing taxonomies and ontologies is rather costly since a lot of manual operations are required. A number of studies proposed automated building of taxonomies based on linguistic resources and/or statistical machine learning, including multiagent settings (Kerschber 2003, Liu & Birnbaum 2007, Curtis et al 2009, Domingos and Poon 2009, Kozareva et al 2009), has been proposed. However, most of these approaches have not found practical applications due to insufficient accuracy of resultant search, limited expressiveness of representations of queries of real users, or high cost associated with manual construction of linguistic resources and their limited adjustability.

Search results snippets is a special case of microtext, where it is hard to apply full-scale linguistic processing since sentences are incomplete. One of the microtext processing task is to refine search relevance by matching queries with search results snippets to filter out irrelevant ones, which is done using syntactic generalization (Galitsky et al 2010), augmented by an ontology-based method. Syntactic generalization is based on finding a set of maximal common sub-trees for a pair of syntactic parse trees for two sentences.

In this paper we also focus on such case of microtext as captions of images and videos as given by authors. To perform mining for videos and images to aggregate them with an article, one needs to match such captions, which are not well structured, with texts. This study draws a commercial evaluation of ontology-based microtext aggregation technology.

In this study we propose automated ontology building mechanism which is based on initial set of main entities (a seed) for given vertical knowledge domain of microtexts. This seed then automatically extended by mining of web documents which include a meaning of a current ontology node. This node is further extended by entities which are the results of inductive learning of commonalities between these documents. These commonalities are extracted using an operation of syntactic generalization, which finds the common parts of syntactic parse trees of a set of documents, obtained for the current ontology node. Syntactic generalization has been extensively evaluated commercially to improve text relevance (Galitsky et al 2010), and in this study we apply it for automated building of ontologies for processing microtexts.

The value of semantically-enabling search engines for improving search relevance has been well understood by the commercial search engine community. Once an 'ideal' ontology is available, properly covering all important entities in a vertical domain, it can be directly applied to filtering out irrelevant answers. The state of the art in this area is how to apply a real-world ontology, far from being ideal and complete, which is automatically learned via web mining, to matching microtexts and to search in general. It is currently clear that lightweight keyword based approaches cannot adequately tackle this problem. In this paper we address it combining web mining as a source of learning, and syntactic generalization as a learning tool.

Improving microtext match relevance by learned ontology

To answer questions and match microtexts, both in natural language and in a keyword-based form, it is beneficial to 'understand' what is question about. In our case this 'understanding' is a preferential treatment of keywords:

In a query with keywords {*a b c*} we understand that query is about *b*, if queries {*a b*} and {*b c*} are marginally relevant, and {*a c*} is irrelevant.

Our narrow notion of query understanding is the ability to say which keywords in the query are essential (such as *b* in the above example), so that without them the other query terms become meaningless, and an answer which does not contain b is irrelevant to the query which includes *b* .

The property of being an essential keyword in a query is hierarchical, for query {*a b c d*}, if *b* is essential, *c* can also be essential when *b* is in the query such that {*a b c*}, {*b c d*} , {*b c*} are relevant, even {*a b*}, {*b d*} are marginally relevant, but {*a d*} is not. Subsets of relevant keywords in a query form a lattice; logical properties of sets of keywords, and logical forms expressing meanings of queries are explored in (Galitsky 2003). There is a systematic way to treat relative importance of keywords via default reasoning (Galitsky 2005); multiple meanings of keyword combinations are represented via operational semantics of default logic.

Taxonomies are required to support query understanding. Taxonomies facilitate the assessments of whether a particular match between query and the answer is relevant or not, based on the above notion of query understanding. Hence for a query {*a b c d*} and two answers (snippets) {*b c d, ... e f g*} and {*a c d ... e f g*} , the former is relevant and the latter is not.

Achieving relevancy using ontology is based on totally different mechanism than a conventional TF*IDF based search. In the latter, importance of terms is based on the frequency of occurrence, and any term can be omitted in the search result if the rest of terms give acceptable relevancy score. In the ontology based search we know which terms *should* occur in the answer and which terms *must* occur there, otherwise the search result becomes irrelevant.

Building ontology by web mining

Our main hypotheses for automated learning taxonomies on the web is that common expressions between search results for given set of entities gives us *parameters* of these entities. Formation of the ontology follows the unsupervised learning style, once the seed terms are fixed. It can be viewed as a human development process, where a baby explores new environment and forms new rules. Initial set of rules is set genetically, and the learning process adjusts these rules to particular habituation environment, to make these rules more sensitive (and therefore allows more beneficial decision making). As new rules are being accepted or rejected during their application process, exposure to new environment facilitates formation of new specific rules. After the new, more complex rules are evaluated and some part of these newly formed rules is accepted, complexity of rules grows further to adapt to further peculiarities of environment.

We learn new entities to extend our ontology in a similar unsupervised learning setting. We start with the seed ontology, which enumerates the main entities of a given domain, and relations of these entities with a few domain-determining concepts. For example, a seed for tax domain will include the relationships

tax - deduct
tax-on-income
tax-on-property,

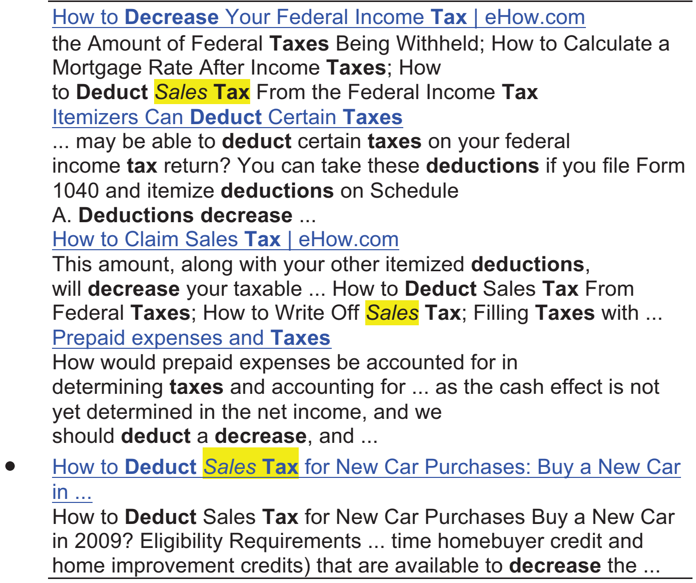
where *tax* is a domain-determining entity, and {*deduct, income, property*} are main entities in this domain. The objective of ontology learning is to acquire further parameters of existing entities such as *tax - deduct*. In the next iteration of learning these parameters will be turned into entities, so that a new set of parameters will be learned.

Learning iteration is based on web mining. To find parameters for given set of tree leaves (current entities), we go to the web and search for common expressions between search results (snippets) for query formed for current tree paths. For the example above, we search for *tax-deduct, tax-on-income, tax-on-property* and extract words and expressions which are **common** between search results. Common words are single verbs, nouns, adjectives and even adverbs or multi-words, including propositional, noun and verb phrases, which occur in **multiple** search results. The central part of our paper, Section 3, explains how to extract common expressions between search results and form new set of current entities (ontology leaves).

After such common words and multi-words are identified, they are added to the original words. For

example, for the path *tax - deduct* newly leaned entities can be *tax-deduct → decrease-by*
tax-deduct → of-income
tax-deduct → property-of
tax-deduct → business
tax-deduct → medical-expense.

Now from the path in the ontology tree *tax – deduct* we obtained five new respective paths. The next step is to collect parameters for each path in the new set of leaves for the ontology tree. In our example, we run five queries and extract parameters for each of them. The results will look like: *tax- deduct-decrease-by → sales*
tax-deduct-decrease-by →401-K
tax-deduct-decrease → medical
tax - deduct- of-income → rental
tax – deduct - of-income → itemized



How to **Decrease** Your Federal Income **Tax** | eHow.com
the Amount of Federal **Taxes** Being Withheld; How to Calculate a Mortgage Rate After Income **Taxes**; How to **Deduct** *Sales* **Tax** From the Federal Income **Tax**

Itemizers Can **Deduct** Certain **Taxes**
... may be able to **deduct** certain **taxes** on your federal income **tax** return? You can take these **deductions** if you file Form 1040 and itemize **deductions** on Schedule A. **Deductions decrease** ...

How to Claim Sales **Tax** | eHow.com
This amount, along with your other itemized **deductions**, will **decrease** your taxable ... How to **Deduct** Sales **Tax** From Federal **Taxes**; How to Write Off *Sales* **Tax**; Filling **Taxes** with ...

Prepaid expenses and **Taxes**
How would prepaid expenses be accounted for in determining **taxes** and accounting for ... as the cash effect is not yet determined in the net income, and we should **deduct** a **decrease**, and ...

- How to **Deduct** *Sales* **Tax** for New Car Purchases: Buy a New Car in ...
How to **Deduct** Sales **Tax** for New Car Purchases Buy a New Car in 2009? Eligibility Requirements ... time homebuyer credit and home improvement credits) that are available to **decrease** the ...

Fig.1: Search results on Bing for the current ontology tree path *tax-deduct-decrease.*

For example, searching the web for *tax-deduct-decrease* allows discovery of an entity *sales-tax* associated with decrease of tax deduction, usually with meaning 'sales tax' (italicized and highlighted in Fig.1). Commonality between snippets shows the sales tax should be taken into account while calculating *tax deduction*, and not doing that would *decrease* it.

Hence the ontology is built via inductive learning of web search results in iterative mode. We start with the ontology seed nodes, then find web search results for all currently available graph paths, and then for each commonality found in these search results we augment each of these ontology paths by adding respective leaf nodes. In other words, for each iteration we discover the list of parameters for each set of currently available entities, and then turn these parameters into entities for the next iteration (Fig.2).

The ontology seed is formed manually or can be compiled from available domain-specific resources. Seed ontology should contain at least 2-3 nodes so that ontology growth process has a meaningful start. Ontology seed can include, for example, a glossary of particular knowledge domain, readily available for a given vertical domain.

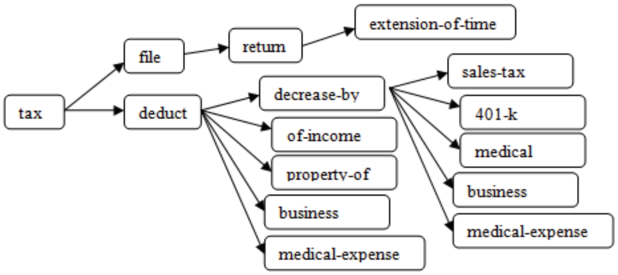


Fig.2 Ontology for tax domain

Filtering answers based on ontology

To use the ontology to filter out irrelevant questions, we search for ontology path closest to the given question in terms of the number of entities from this question. Then this path and leave node specify most accurate meaning of the question, and constrain which entities *must* occur and which *should* occur in the answer to be considered relevant. If the n-th node entity from the question occurs in answer, then all k < n entities should occur in it as well. In-depth treatment of this property is presented as a default reasoning framework in (Galitsky 2005).

Examples above illustrate this main requirement. Naturally, multiple ontology paths can be similar to the question, then the above should hold for at least one of these paths. Ontologiesies help to solve disambiguation problem: for a question

(Q) "When can I file extension of time for my tax return?"

let us imagine two answers:

(A1) "You need to file form 1234 to request a 4 month extension of time to file your tax return"

(A2) "You need to download file with extension 'pdf', print and complete it to file your tax return".

We expect the closest ontology path to be :

(T) tax - file-return - extension-of-time.

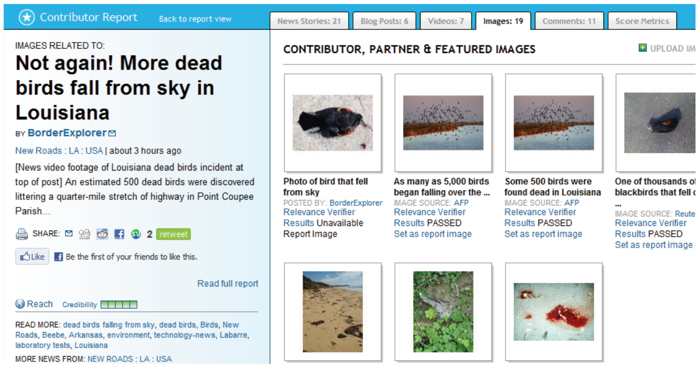
Here *tax* is a main entity, *file-return* we expect to be in the seed, and *extension-of-time* would be the learned entity, so A1 will match with ontology and is an acceptable answer, and A2 is not.

Evaluation of microtext similarity improvement

We subject the proposed technique of ontology-based and syntactic generalization-based techniques in the commercial main of news analysis at AllVoices.com. The task is to cluster relevant news together, by means of text relevance analysis. By definition, multiple news articles belong to the same cluster, if there is a substantial overlap of involved entities such as geo locations and names of individuals, organizations and other agents, as well as relations between them. Some of these can be extracted by entity taggers, and/or by using taxonomies, and some are handled in real time using syntactic generalization. The latter is applicable if there is a lack of prior entity information.

In addition to forming a cluster of relevant documents (Fig. 4), it is necessary to aggregate relevant images and videos from different sources such as Google image, YouTube and Flickr, and access their relevance given their textual descriptions and tags, where the similar ontology and syntactic generalization-based technique is applied (Fig. 3, Table 2).

Precision of text analysis is achieved by site usability (click rate) by more than nine million unique visitors per month. Recall is accessed manually; however the system needs to find at least a few articles, images and videos for each incoming article. Usually, for web mining and web document analysis recall is not an issue, it is assumed that there is a high number of articles, images and videos on the web for mining.



Contributor Report Back to report view

News Stories: 21 | Blog Posts: 6 | Videos: 7 | Images: 19 | Comments: 11 | Score Metrics

IMAGES RELATED TO:

Not again! More dead birds fall from sky in Louisiana

BY BorderExplorer

New Roads : LA : USA | about 3 hours ago

[News video footage of Louisiana dead birds incident at top of post] An estimated 500 dead birds were discovered littering a quarter-mile stretch of highway in Point Coupee Parish...

SHARE: 2 retweet

Like Be the first of your friends to like this.

Read full report

Reach Credibility

READ MORE: dead birds falling from sky, dead birds, Birds, New Roads, Beebe, Arkansas, environment, technology-news, Labarre, laboratory tests, Louisiana

MORE NEWS FROM: NEW ROADS : LA : USA

CONTRIBUTOR, PARTNER & FEATURED IMAGES

UPLOAD IM

Photo of bird that fell from sky
POSTED BY: BorderExplorer
Relevance Verifier Results Unavailable
Report Image

As many as 5,000 birds began falling over the ...
IMAGE SOURCE: AFP
Relevance Verifier Results PASSED
Set as report image

Some 500 birds were found dead in Louisiana
IMAGE SOURCE: AFP
Relevance Verifier Results PASSED
Set as report image

One of thousands of blackbirds that fell ...
IMAGE SOURCE: Reute
Relevance Verifier Results PASSED
Set as report image

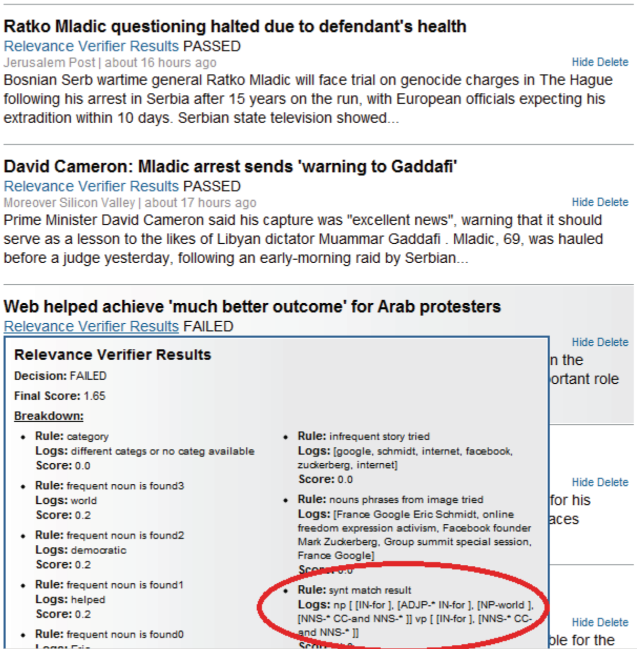
Fig.3: A news articles and aggregated images found on the web and determined to be relevant to this article.

Precision data for the relevance relation between an articles and other article, blog posting, image and vides is presented in Table 2 (normalized taking into account the decreased recall). Notice that although the ontology-based method on its own, without other relevance means, has a very low precision and does not outperform the baseline of the statistical assessment, there is a noticeable improvement of precision in hybrid system. We can conclude that syntactic generalization and ontology-based methods (which also rely on syntactic generalization) use different sources of relevance information, so they are indeed complementary to each other.

The objective of syntactic generalization was to filter out false-positive relevance decision, made by statistical relevance engine designed following (Liu & Birnbaum 2007, Liu & Birnbaum 2008). The percentage of false-positive news stories was reduced from 29 to 13 (about 30000 stories/month viewed by 9 million unique users), and the percentage of false positive image attachment was reduced from 24 to 18 (about 3000 images and 500 videos attached to stories monthly).

Table 2: Improvement the precision of text similarity

Media/ method of text similarity assessment	Full size news articles	Abstracts of articles	Blog posting	Comments	Images	Videos
Frequencies of terms in documents	29.3%	26.1%	31.4%	32.0%	24.1%	25.2%
Syntactic generalization	17.8%	18.4%	20.8%	27.1%	20.1%	19.0%
Ontology-based	45.0%	41.7%	44.9%	52.3%	44.8%	43.1%
Hybrid (ontology + syntactic)	13.2%	13.6%	15.5%	22.1%	18.2%	18.0%



Ratko Mladic questioning halted due to defendant's health
Relevance Verifier Results PASSED
Jerusalem Post | about 16 hours ago
Hide Delete
Bosnian Serb wartime general Ratko Mladic will face trial on genocide charges in The Hague following his arrest in Serbia after 15 years on the run, with European officials expecting his extradition within 10 days. Serbian state television showed...

David Cameron: Mladic arrest sends 'warning to Gaddafi'
Relevance Verifier Results PASSED
Moreover Silicon Valley | about 17 hours ago
Hide Delete
Prime Minister David Cameron said his capture was "excellent news", warning that it should serve as a lesson to the likes of Libyan dictator Muammar Gaddafi . Mladic, 69, was hauled before a judge yesterday, following an early-morning raid by Serbian...

Web helped achieve 'much better outcome' for Arab protesters
Relevance Verifier Results FAILED
Hide Delete

Relevance Verifier Results
Decision: FAILED
Final Score: 1.65
Breakdown:

- **Rule:** category
  Logs: different categs or no categ available
  Score: 0.0
- **Rule:** frequent noun is found3
  Logs: world
  Score: 0.2
- **Rule:** frequent noun is found2
  Logs: democratic
  Score: 0.2
- **Rule:** frequent noun is found1
  Logs: helped
  Score: 0.2
- **Rule:** frequent noun is found0
- **Rule:** infrequent story tried
  Logs: [google, schmidt, internet, facebook, zuckerberg, internet]
  Score: 0.0
- **Rule:** nouns phrases from image tried
  Logs: [France Google Eric Schmidt, online freedom expression activism, Facebook founder Mark Zuckerberg, Group summit special session, France Google]
- **Rule:** synt match result
  Logs: np [[IN-for], [ADJP-* IN-for], [NP-world], [NNS-* CC-and NNS-*]] vp [[IN-for], [NNS-* CC-and NNS-*]]

n the
ortant role
Hide Delete
for his
aces
Hide Delete
ble for the

Fig.4: Blog clustering example. Encoding for the syntactic generalization between article title and blog message title is shown in oval.

Conclusions: ontologies and syntactic generalizations for microtext

We believe that ontologies and more sensitive match of keywords (compared to bag-of-words and TF*IDF) are the means for microtext processing. Since microtext is muddled with abbreviations and acronyms, and we don't 'know' all mappings, semantic analysis should be tolerant to omits of some entities and still understand "what this text fragment is about". Since we are unable to filter out noise “statistically” like most NLP environments do, we have to rely on ontology. Syntactic generalization takes bag-of-words and pattern-matching classes of approaches to the next level allowing to treat unknown words systematically as long as their part of speech information is available from context.

Working definition of microtext according to (Ellen 2011) includes a single author contribution which is rather brief, using informal and unconventional grammar, and combination of meta-data and conventional free-text. Subjects of this study such as search results snippets, and especially image and video captions fit this definition. What we have not tackled in this study is conversational threading. However we believe the technique developed for the subjects of this study can be naturally extended towards other forms of microtext including instant messages, chat rooms, transcribed voice communications, and microblog services.

For a few decades, most approaches to NL semantics relied on mapping to First Order Logic representations with a general prover and without using acquired rich knowledge sources. Significant development in NLP, specifically the ability to acquire knowledge and induce some level of abstract representation such as taxonomies is expected to support more sophisticated and robust approaches. A number of recent approaches are based on shallow representations of the text that capture lexico-syntactic relations based on dependency structures and are mostly built from grammatical functions extending keyword matching (Durme et al 2003). On the contrary, ontology learning in this work is performed in a vertical domain, where ambiguity of terms is limited, and therefore fully automated settings produce adequate resultant search accuracy. Hence our approach is finding a number of commercial applications including relevancy engine at citizens’ journalism portal AllVoices.com and search and recommendation at Zvents.com. The writing style of non-professional / citizens’ journalism such as AllVoices, and especially their discussion threads, fit well the general notion of microtext

Usually, classical approaches to semantic inference rely on complex logical representations. However, practical applications usually adopt shallower lexical or lexical-syntactic representations, but lack a principled inference framework. The current work deals with syntactic tree transformation in the graph learning framework (compare with Chakrabarti & Faloutsos 2006), treating various phrasings for the same meaning in a more unified and automated manner.

Traditionally, semantic parsers are constructed manually, or are based on manu-ally constructed semantic ontologies, but these are is too delicate and costly. A num-ber of supervised learning approaches to building formal semantic representation have been proposed (Cardie & Mooney 2007). Unsupervised approaches have been proposed as well, however they applied to shallow semantic tasks (e.g., paraphrasing (Lin and Pantel, 2001), information extraction (Banko et al., 2007), and semantic parsing (Poon and Domingos 2008). The problem domain in the current study required much deeper handling of syntactic peculiarities to build taxonomies. In terms of learning, our approach is closer in merits to unsupervised learning of complete formal semantic representation. Compared to semantic role labeling (Carreras and Marquez, 2004) and other forms of shallow semantic processing, our approach maps text to formal meaning representations, obtained via generalization.

The study of concepts can advance further by clarifying the meanings of basic terms such as “prototype” and by constructing a large-scale primary ontology of concept types (Howard 1992). Based on concept structures, two secondary concept taxonomies and one of conceptual structures has been built, where the primary ontology organizes much data and several previous taxonomies into a single framework. It suggests that many concept types exist, and that type determines how a concept is learned, is used and how it develops. (Alany & Brewster 2005) provides a tool to facilitate the re-use of existing knowledge structures such as taxonomies, based on the ranking of ontologies. This tool uses as input the search terms provided by a knowledge engineer and, using the output of an ontology search engine, ranks the taxonomies. A number of metrics in an attempt to investigate their appropriateness for ranking ontologies has been applied, and results were compared with a questionnaire-based human study.

(Carlson et al 2010) addressed problem of building an intelligent computer agent that runs forever and that each day must extract information from the web to populate a growing structured knowledge base, and also learn to perform this task better than on the previous day. A set of design principles are proposed for such an agent, a partial implementation of such a system is described that has already learned to extract a knowledge base containing over quarter of a million beliefs with an estimated precision of 74%. (Balakrishna et al 2010) presents a generalized and improved procedure to automatically extract deep semantic information from text resources and rapidly create semantically-rich domain ontologies while keeping the manual intervention to a minimum. The authors present evaluation results for the intelligence and financial ontology libraries, semi-automatically created by their proposed methodologies using freely-available textual resources from the Web. Although the experimental evaluations in these studies are impressive, an industrial type of evaluation would be useful to compare them with the current study.

Use of syntactic generalization in this work is two-fold. Firstly, it is used off-line to form the node of ontology tree, finding commonalities between search results for a given ontology node. Secondly, syntactic generalization is used online for measuring similarity of either two portions of text, or question and answer, to measure the relevance between them. We demonstrated that merging ontology-based methods and syntactic generalization methods improves the relevance of text understanding in general, and complementary to each other, because the former uses pure meaning-based information , and the latter user linguistic information about the involved entities. Naturally, such combination outperforms a bag-of-words approach in horizontal domain, and also, according to our evaluation, outperforms a baseline statistical approach in vertical domains.

References

Alani, H. and Brewster, C. Ontology ranking based on the analysis of concept structures. K-CAP '05 Proceedings of the 3rd international conference on Knowledge capture 2005.

Allen, J.F. Natural Language Understanding, Benjamin Cummings, 1987.

Banko, M., Cafarella, J., Soderland, S., Broadhead M., and Etzioni O.. 2007 Open information extraction from the web. In *Proceedingsof the Twentieth International Joint Conference on Artificial Intelligence*, pages 2670–2676, Hyderabad, India. AAAI Press.

Chakrabarti, D. and C. Faloutsos, “Graph Mining: Laws, Generators, and Algorithms,” *ACM Computing Surveys,* vol. 38, no. 1, 2006

Cardie, C., Mooney R.J. Machine Learning and Natural Language. Machine Learning 1(5) 1999.

Carreras X. and Luis Marquez. 2004. Introduction to the CoNLL-2004 shared task: Semantic role labeling. In *Proceedings of the Eighth Conference on Computational Natural Language Learning*, pp 89–97, Boston, MA. ACL.

Ellen, J. (2011). All about microtext: A working definition and a survey of current microtext research within artificial intelligence and natural language processing. In Proceedings of the Third International Conference on Agents and Artificial Intelligence. Rome, Italy: Springer.

Galitsky, B. Natural Language Question Answering System: Technique of Semantic Headers. Advanced Knowledge International, Australia 2003.

Galitsky, B., Dobrocsi, G., de la Rosa, JL, Kuznetsov SO: From Generalization of Syntactic Parse Trees to Conceptual Graphs. ICCS 2010: 185-190.

Galitsky, B. Disambiguation Via Default Rules Under Answering Complex Questions. Intl J AI Tools 14(1-2), World Scientific, 2005.

Howard, R.W. Classifying types of concept and conceptual structure: Some taxonomies. Journal of Cognitive Psychology, V4, Issue 2 April 1992, 81 - 111.

Robinson JA. A machine-oriented logic based on the resolution principle. *Journal of the Association for Computing Machinery*, 12:23-41, 1965

Lin, D., and Pantel, P. 2001. DIRT: discovery of inference rules from text. In *Proc. of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2001*, 323–328.

Domingos P. and Poon, H. Unsupervised Semantic Parsing, with, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 2009. Singapore: ACL.

Durme, B. V.; Huang, Y.; Kupsc, A.; and Nyberg, E. 2003. Towards light semantic processing for question answering. HLT Workshop on Text Meaning.

Kozareva, Z., Eduard Hovy, and Ellen Riloff , Learning and Evaluating the Content and Structure of a Term Ontology. Learning by Reading and Learning to Read AAAI Spring Symposium 2009. Stanford CA.

Liu, J. Larry Birnbaum, 2008. What do they think?: Aggregating local views about news events and topics. WWW: 1021-1022.

Liu, J., Birnbaum, L 2007. Measuring Semantic Similarity between Named Entities by Searching the Web Directory. Web Intelligence: 461-465.

Kerschberg, L W. Kim, and A. Scime, 2003. A Semantic Ontology-Based Personalizable Meta-Search Agent," in Innovative Concepts for Agent-Based Systems, vol. LNAI 2564, Lecture Notes in Artificial Intelligence, W. Truszkowski, Ed. Heidelberg: Springer-Verlag, pp. 3-31.

Zhang, M., Zhou GD, Aw, A. Exploring syntactic structured features over parse trees for relation extraction using kernel methods, Information Processing and Management: an International Journal V 44 , Issue 2 (March 2008) 687-701.

Curtis, J., David Baxter, Peter Wagner, John Cabral, Dave Schneider, Michael Witbrock. Methods of Rule Acquisition in the TextLearner System. Learning by Reading and Learning to Read AAAI Spring Symposium 2009. Stanford CA.

Carlson, A., J. Betteridge, B. Kisiel, B. Settles, E.R. Hruschka Jr. and T.M. Mitchell. 2010. Toward an Architecture for Never-Ending Language Learning. (AAAI-2010).

Balakrishna,M., Dan Moldovan, Marta Tatu, Marian Olteanu. 2010. Semi-Automatic Domain Ontology Creation from Text Resources. International Conference on Language Resources and Evaluation.