# Using Generalization of Syntactic Parse Trees for Taxonomy Capture on the Web

Boris A. Galitsky[1], Gábor Dobrocsi[1], Josep Lluis de la Rosa[1], and Sergei O. Kuznetsov[2]

[1] University of Girona, Girona, Catalonia, Spain
bgalitsky@hotmail.com, gadomail@gmail.com, peplluis@silver.udg.edu
[2] Higher School of Economics, Moscow Russia
skuznetsov@yandex.ru

**Abstract.** We implement a scalable mechanism to build a taxonomy of entities which improves  relevance of search engine in a vertical domain. Taxonomy 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. Taxonomy and syntactic generalization is applied to relevance improvement in search and text similarity assessment in commercial setting; evaluation results show substantial contribution of both sources.

**Keywords:** learning taxonomy, learning syntactic parse tree, syntactic generalization, search relevance.

## 1 Introduction

Nowadays, designing search engines and text relevance systems, it is hard to overestimate the role of taxonomies for improving their precisions, especially in vertical domains. However, 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 [19, 21, 22].  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.

In this study we propose automated taxonomy building mechanism which is based on initial set of main entities (a seed) for given vertical knowledge domain. This seed is then automatically extended by mining of web documents which include a meaning of a current taxonomy 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 taxonomy node.

S. Andrews et al. (Eds.): ICCS 2011, LNAI 6828, pp. 104–117, 2011.

Syntactic generalization has been extensively evaluated commercially to improve text relevance [8, 9], and in this study we apply it for automated building of taxonomies.

Proceeding from parsing to semantic level is an important task towards natural language understanding, and has immediate applications in tasks such as information extraction and question answering [3, 5, 13]. In the last ten years there has been a dramatic shift in computational linguistics from manually constructing grammars and knowledge bases to partially or totally automating this process by using statistical learning methods trained on large annotated or non-annotated natural language corpora. However, instead of using such corpora, in this paper we use web search results for common queries, since their accuracy is higher and they are more up-to-date than academic linguistic resources.

The value of semantically-enabling search engines for improving search relevance has been well understood by the commercial search engine community [2]. Once an ‘ideal’ taxonomy 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 taxonomy to search relevance improvement, where such a taxonomy is automatically compiled from the web and therefore is far from being ideal. It has become obvious that lightweight keyword based approaches cannot adequately tackle this problem. In this paper we address it combining web mining as a source of training set, and syntactic generalization as a learning tool.

## 2 Improving Search Relevance by Ontologies

To answer a question, natural language or keyword-based, it is beneficial to 'understand' what is this question about. In the sense of current paper this 'understanding' is a preferential treatment of keywords. We use the following definition of a relationship between a set of keywords and its element *is-about* (*set-of-keywords, keyword).*

For a query with keywords {*a b c*} we understand that query is about *b*, if queries {*a b*} and {*b c*} are relevant or marginally relevant, and {*a c*} is irrelevant. Our definition of query understanding, which is rather narrow, 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* .

For example, in the set of keywords {*computer, vision, technology*}, {*computer, vision*}, {*vision, technology*} are relevant, and {computer, technology} are not, so the query *is about* vision. Notice that if a set of keywords form a noun phrase or a verb phrase, it does not necessarily mean that the head or a verb is a keyword this ordered set is about. Also notice that we can group words into phrases when they form an entity:

*is-about*({*vision, bill, gates*}, Ø) , whereas
*is-about*({*vision, bill-gates, in- computing},* bill-gates).

We refer to a keyword as *essential* if it occurs on the right side of *is-about.*

To properly formalize the latter observation, we generalize *is-about* relations towards the relation between a set of keywords and its subset. For query {*a b c d*}, if *b* is essential (*is-about*({*a b c d*}, {*b*}), *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 (*is-about*({*a b c d*}, {*b,c*}). Logical properties of sets of keywords, and logical forms expressing meanings of queries, are explored in [8]. There is a systematic way to treat relative importance of keywords via default reasoning [10]; 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 a query and an answer is relevant or not, based on the above notion of query understanding via *is-about* relation. 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 a taxonomy 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 taxonomy-based search we know which terms *should* occur in the answer and which terms *must* occur there, otherwise the search result becomes irrelevant.

## 2.1 Building Taxonomy by Web Mining

Our main hypotheses for automated learning taxonomies on the web is that common expressions between search results for a given set of entities gives us *parameters* of these entities. Formation of the taxonomy follows the unsupervised learning style. 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 taxonomy in a similar unsupervised learning setting. We start with the seed taxonomy, 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 taxonomy 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 (Fig. 1).

Learning iteration is based on web mining. To find parameters for a given set of tree leaves (current entities), we go to the web and search for common expressions between the 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 (taxonomy leaves).

After such common words and multi-words are identified, they are added to the original words. E.g. for the path *tax - deduct* newly learned entities can be

*tax-deduct → decrease-by* *tax-deduct → of-income*
*tax-deduct → property-of* *tax-deduct → business*
*tax-deduct → medical-expense.*

The format here is *existing_entity → its parameter (to become a new_entity), '→'* here is an unlabeled ontology edge.

Now from the path in the taxonomy 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 taxonomy tree. In our example, we run five queries and extract parameters for each of them. The results will look like:

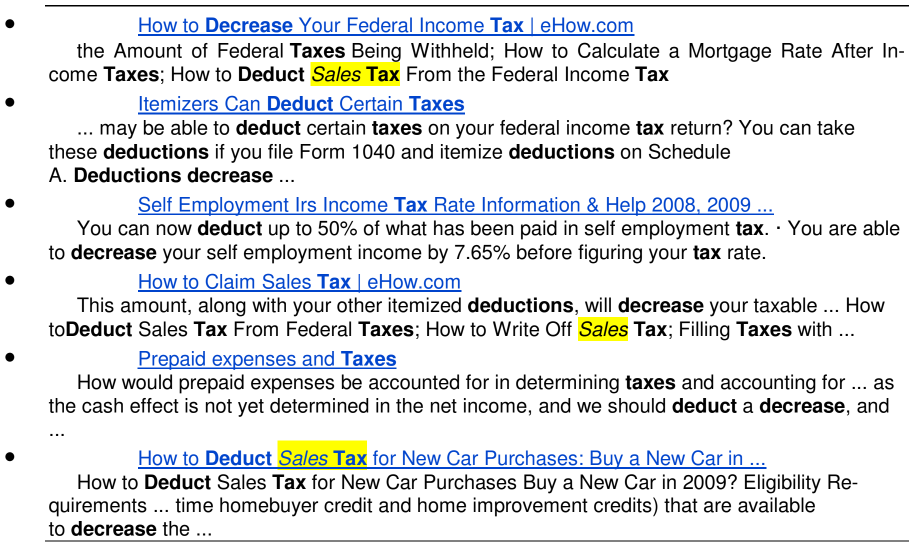
*tax- deduct-decrease-by → sales*
*tax-deduct-decrease-by →401-K*
*tax-deduct-decrease → medical*

*tax - deduct- of-income → rental*
*tax – deduct - of-income → itemized*
*tax – deduct – of-income → mutual-funds*

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 taxonomy is built via inductive learning of web search results in iterative mode. We start with the taxonomy 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 taxonomy 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 taxonomy seed is formed manually or can be compiled from available domain-specific resources. Seed taxonomy should contain at least 2-3 nodes so that taxonomy growth process has a meaningful start. Taxonomy seed can include, for example, a glossary of particular knowledge domain, readily available for a given vertical domain, like http://www.investopedia.com/categories/taxes.asp for tax entities.



- 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** ...
- Self Employment Irs Income **Tax** Rate Information & Help 2008, 2009 ...
  You can now **deduct** up to 50% of what has been paid in self employment **tax**. · You are able to **decrease** your self employment income by 7.65% before figuring your **tax** rate.
- 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.com for the current taxonomy tree path *tax-deduct-decrease*

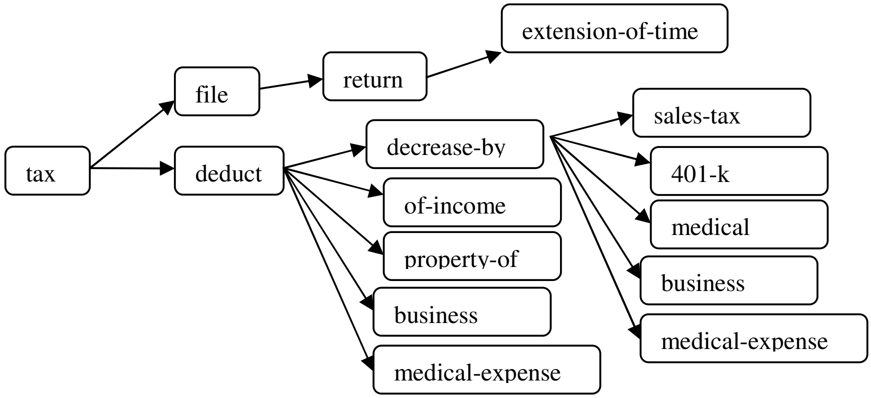


**Fig. 2.** Taxonomy for tax domain

## 2.2 Filtering Answers Based on Taxonomy

To use the taxonomy to filter out irrelevant questions, we search for taxonomy path (down to a leaf node if possible) which is 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.

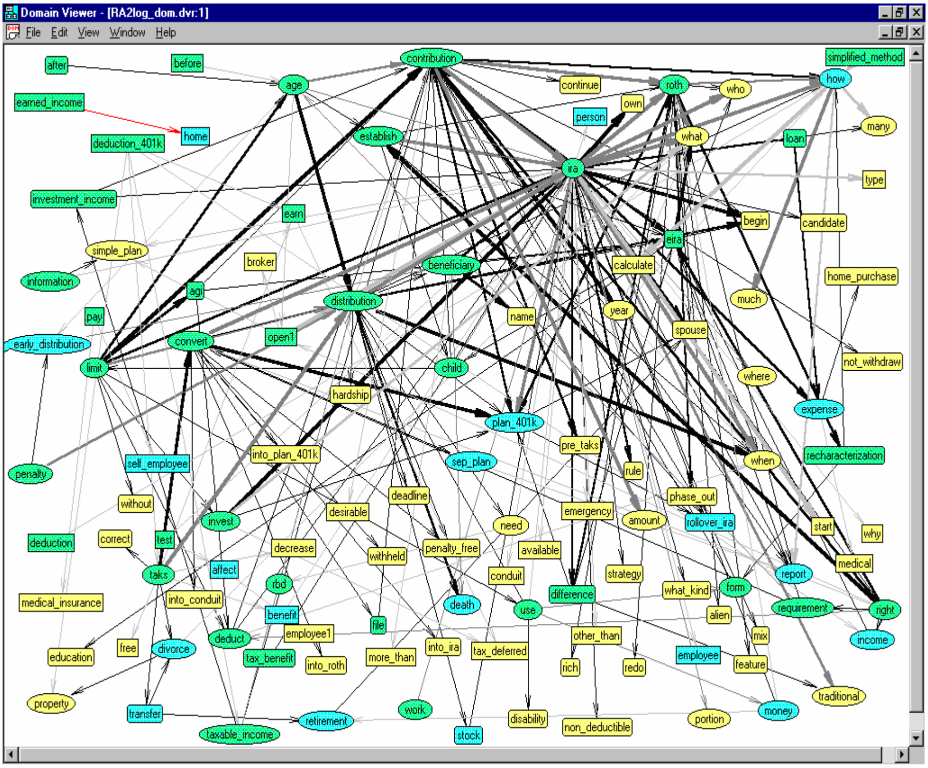


**Fig. 3.** Visualization of a log graph

For the majority of search applications, *acceptability* condition is easier to apply that the 'most accurate' condition: An answer $A$ is acceptable if it includes all essential keywords from the question $Q$ as found in the taxonomy path $T_p \in T$

$A \subseteq T_p \cap Q$ .

For the best answer we write $A_{best}$ : max(cardinality $(A_{best} \cap( T_p \cap Q))$ ,

where S ∩ G is an operation of finding a maximal path in a graph G whose node labels belong to a set S, so that *is-about*(*K*, *S*) for some *K*.

Examples above illustrate this main requirement. Naturally, multiple taxonomy these paths. Taxonomies 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 taxonomy path to be :

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

*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 taxonomy and is an acceptable answer, and A2 is not.

Another way to represent taxonomy is not to enforce it to be a tree, **least general** but only allow one node for each label instead (Fig 3).

## 3 Syntactic Generalization of Sentences

To measure similarity of abstract entities expressed by logic formulas, the least-general generalization (also called anti-unification) was proposed for a number of machine learning approaches, including explanation-based learning and inductive logic programming. It is the opposite of most general unification, therefore it is also called anti-unification [12]. To form commonality expression between search result snippets (in general, to measure similarity between 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 of logical formulas would be sufficient. However, in text mining problem, we need to deal with original text, so we apply generalization to syntactic parse trees obtained for each sentence. Rather than extracting common keywords, generalization operation produces a syntactic expression that can be semantically interpreted as a common meaning shared by two sentences, which may underlie a new entity for taxonomy node.

1) Obtain parsing tree for each sentence. For each word (tree node) we have <lemma, part of speech , word form> information. This information is contained in the node label. We also have an arc to the other node.
2) Split sentences into sub-trees which are phrases of 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. Generalize each pair of sub-trees for both sentences for each phrase type.
5) For each pair of sub-trees perform an alignment of phrases, and then generalize each node of these aligned sentences as sub-trees. For the obtained set of trees (generalization results), calculate the score which is a POS-weighted sum of the number of nodes for all trees from this set (see details in [9]).
6) For each pair of sub-trees for phrases, select the set of generalizations with highest score (least general).
7) Form the sets of generalizations for each phrase types whose elements are sets of generalizations for this type.
8) Filtering the list of generalization results: for the list of generalization for each phrase type, select least general elements from this list of generalization for a given pair of phrases.

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 results are the sets of sub-trees {*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.

The result of generalization can be further generalized with other parse trees or generalizations. For a set of sentences, the 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. Further details on syntactic generalization can be obtained in [9].

## 4 Evaluation of Search Relevance Improvement

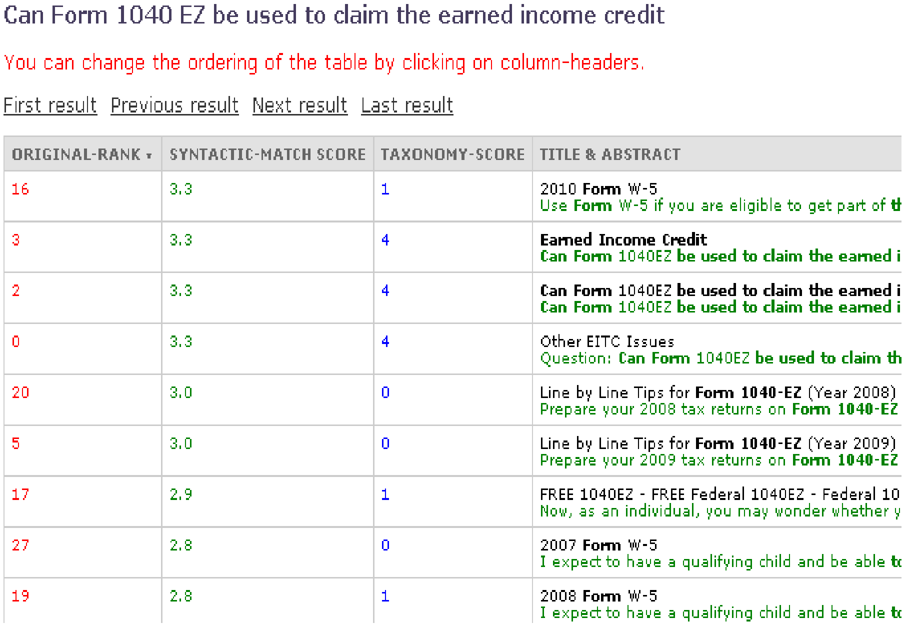
Evaluation of search included an assessment of classification accuracy for search results as relevant and irrelevant. Since we used the generalization score between the query and each hit snapshot, we drew a threshold of five highest score results as relevant class and the rest of search results as irrelevant. We used the Yahoo search API and applied the generalization score to find the highest score hits from first fifty Yahoo search results (Fig. 4). We then consider the first five hits with the highest generalization score (not Yahoo score) to belong to the class of relevant answers. Third and second rows from the bottom contain classification results for the queries of 3-4 keywords which is slightly more complex than an average one (3 keywords); and significantly more complex queries of 5-7 keywords, respectively.

The total average accuracy (F-measure) for all above problems is 79.2%. Since the syntactic generalization was the only source of classification, we believe the accuracy is satisfactory. A practical application would usually use a hybrid approach with rules and keyword statistic which would deliver higher overall accuracy, but such application is beyond the scope of this paper. Since the generalization algorithm is deterministic, higher accuracy can be also achieved by extending training set.

In this study we demonstrated that such high-level sentences semantic features as *being informative* can be learned from the low level linguistic data of complete parse tree. Unlike the traditional approaches to *multilevel* derivation of semantics from syntax, we explored the possibility of linking low level but detailed syntactic level with high-level pragmatic and semantic levels *directly*.

**Table 1.** Evaluation of classification accuracy

| Type of search query | Relevancy of Yahoo search, %, averaging over 10 | Relevancy of re-sorting by generalization, %, averaging over 10 | Relevancy compared to baseline, % |
|---|---|---|---|
| 3-4 word phrases | 77 | 77 | 100.0% |
| 5-7 word phrases | 79 | 78 | 98.7% |
| 8-10 word single sentences | 77 | 80 | 103.9% |
| 2 sentences, >8 words total | 77 | 83 | 107.8% |
| 3sentences,>12 words total | 75 | 82 | 109.3% |



Can Form 1040 EZ be used to claim the earned income credit

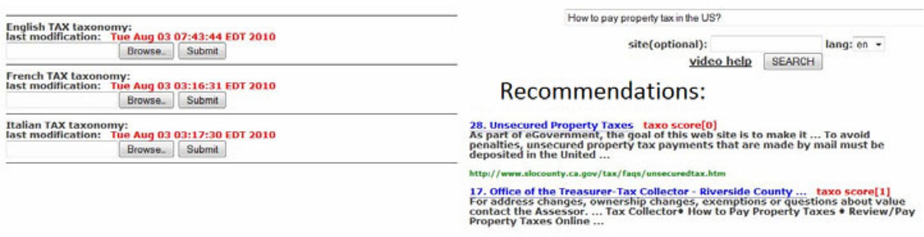
You can change the ordering of the table by clicking on column-headers.

First result Previous result Next result Last result

| ORIGINAL-RANK ▾ | SYNTACTIC-MATCH SCORE | TAXONOMY-SCORE | TITLE & ABSTRACT |
|---|---|---|---|
| 16 | 3.3 | 1 | 2010 **Form** W-5<br>Use **Form** W-5 if you are eligible to get part of **th** |
| 3 | 3.3 | 4 | **Earned Income Credit**<br>**Can Form** 1040EZ **be used to claim the earned i** |
| 2 | 3.3 | 4 | **Can Form** 1040EZ **be used to claim the earned i**<br>**Can Form** 1040EZ **be used to claim the earned i** |
| 0 | 3.3 | 4 | Other EITC Issues<br>Question: **Can Form** 1040EZ **be used to claim th** |
| 20 | 3.0 | 0 | Line by Line Tips for **Form 1040-EZ** (Year 2008)<br>Prepare your 2008 tax returns on **Form 1040-EZ** |
| 5 | 3.0 | 0 | Line by Line Tips for **Form 1040-EZ** (Year 2009)<br>Prepare your 2009 tax returns on **Form 1040-EZ** |
| 17 | 2.9 | 1 | FREE 1040EZ - FREE Federal 1040EZ - Federal 10<br>Now, as an individual, you may wonder whether y |
| 27 | 2.8 | 0 | 2007 **Form** W-5<br>I expect to have a qualifying child and be able **to** |
| 19 | 2.8 | 1 | 2008 **Form** W-5<br>I expect to have a qualifying child and be able **to** |

**Fig. 4.** Sorting search results by syntactic generalization vs taxonomy-based for a given query

We selected Citizens Advise Services as another application domain where taxonomy improves relevance of recommendations easy4.udg.edu/isac/eng/index.php [17,18]. Taxonomy learning of the tax domain was conducted in English and then translated in Spanish, French, German and Italian. It was evaluated by project partners using the tool in Fig 5, where to improve search precision a project partner in a particular location modifies the automatically learned taxonomy to fix a particular case, upload the taxonomy version adjusted for a particular location and verify the improvement of relevance.



English TAX taxonomy:
last modification: Tue Aug 03 07:43:44 EDT 2010

French TAX taxonomy:
last modification: Tue Aug 03 03:16:31 EDT 2010

Italian TAX taxonomy:
last modification: Tue Aug 03 03:17:30 EDT 2010

How to pay property tax in the US?

site(optional): lang: en

Recommendations:

28. Unsecured Property Taxes taxo score[0]
As part of eGovernment, the goal of this web site is to make it ... To avoid penalties, unsecured property tax payments that are made by mail must be deposited in the United ...

http://www.slocounty.ca.gov/tax/faqs/unsecuredtax.htm

17. Office of the Treasurer-Tax Collector - Riverside County ... taxo score[1]
For address changes, ownership changes, exemptions or questions about value contact the Assessor. ... Tax Collector• How to Pay Property Taxes • Review/Pay Property Taxes Online ...

**Fig. 5.** A tool for manual adjustment of taxonomy for providing Citizens Recommendation Services, http://box.cs.rpi.edu:8080/wise/taxo.jsp

## 4.1 Commercial Evaluation of Text Similarity Improvement

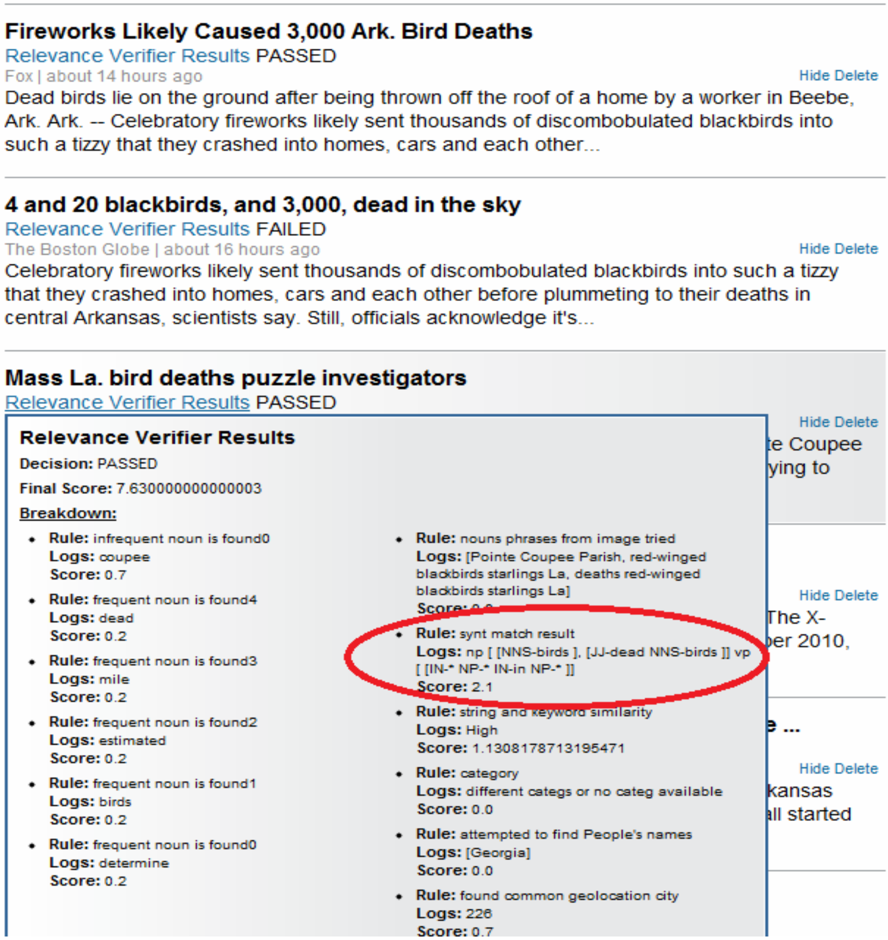
We subject the proposed technique of taxonomy-based and syntactic generalization-based techniques in the commercial area of news analysis at AllVoices.com. The task is to cluster relevant news together, by means of text relevance analysis. By definition, multiple news articles belong to the same cluster, 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 (Fig.7, oval on the right). The latter is applicable if there is a lack of prior entity information.

In addition to forming a cluster of relevant documents, 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 taxonomy and syntactic generalization-based technique is applied (Fig. 6).

Precision of text analysis is achieved by site usability (click rate) of 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 are a high number of articles, images and videos on the web for mining.

Precision data for the relevance relation between an article and other articles, blog postings, images and videos is presented in Table 2 (the percentages are normalized taking into account the decreased recall). Notice that although the taxonomy-based method on its own 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 taxonomy-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 [21,22]. 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).



**Fireworks Likely Caused 3,000 Ark. Bird Deaths**
Relevance Verifier Results PASSED
Fox | about 14 hours ago
Hide Delete
Dead birds lie on the ground after being thrown off the roof of a home by a worker in Beebe, Ark. Ark. -- Celebratory fireworks likely sent thousands of discombobulated blackbirds into such a tizzy that they crashed into homes, cars and each other...

**4 and 20 blackbirds, and 3,000, dead in the sky**
Relevance Verifier Results FAILED
The Boston Globe | about 16 hours ago
Hide Delete
Celebratory fireworks likely sent thousands of discombobulated blackbirds into such a tizzy that they crashed into homes, cars and each other before plummeting to their deaths in central Arkansas, scientists say. Still, officials acknowledge it's...

**Mass La. bird deaths puzzle investigators**
Relevance Verifier Results PASSED
Hide Delete
te Coupee
ying to

**Relevance Verifier Results**

**Decision:** PASSED

**Final Score:** 7.630000000000003

**Breakdown:**

- **Rule:** infrequent noun is found0
  **Logs:** coupee
  **Score:** 0.7
- **Rule:** frequent noun is found4
  **Logs:** dead
  **Score:** 0.2
- **Rule:** frequent noun is found3
  **Logs:** mile
  **Score:** 0.2
- **Rule:** frequent noun is found2
  **Logs:** estimated
  **Score:** 0.2
- **Rule:** frequent noun is found1
  **Logs:** birds
  **Score:** 0.2
- **Rule:** frequent noun is found0
  **Logs:** determine
  **Score:** 0.2
- **Rule:** nouns phrases from image tried
  **Logs:** [Pointe Coupee Parish, red-winged blackbirds starlings La, deaths red-winged blackbirds starlings La]
  **Score:** 0.0
- **Rule:** synt match result
  **Logs:** np [ [NNS-birds ], [JJ-dead NNS-birds ]] vp [ [IN-* NP-* IN-in NP-* ]]
  **Score:** 2.1
- **Rule:** string and keyword similarity
  **Logs:** High
  **Score:** 1.1308178713195471
- **Rule:** category
  **Logs:** different categs or no categ available
  **Score:** 0.0
- **Rule:** attempted to find People's names
  **Logs:** [Georgia]
  **Score:** 0.0
- **Rule:** found common geolocation city
  **Logs:** 226
  **Score:** 0.7

Hide Delete
The X-
ber 2010,

e ...

Hide Delete
kansas
ll started

**Fig. 6.** Exp*lanation for relevance decision while forming a cluster of news articles for the one on Fig.6. The circled area shows the syntactic generalization result for the seed articles and the given one.

**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% |
| Taxonomy-based | 45.0% | 41.7% | 44.9% | 52.3% | 44.8% | 43.1% |
| Hybrid (taxonomy + syntactic) | 13.2% | 13.6% | 15.5% | 22.1% | 18.2% | 18.0% |

## 5 Related Work and Conclusions

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 [15]. On the contrary, taxonomy 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.

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. A generic semantic inference framework that operates directly on syntactic trees has been proposed. New trees are inferred by applying entailment rules, which provide a unified representation for varying types of inferences. Rules are generated by manual and automatic methods, covering generic linguistic structures as well as specific lexical-based inferences. The current work deals with syntactic tree transformation in the graph learning framework (compare with [4, 16]), treating various phrasings for the same meaning in a more unified and automated manner.

Traditionally, semantic parsers are constructed manually, or are based on manually constructed semantic ontologies, but these are too delicate and costly. A number of supervised learning approaches to building formal semantic representation have been proposed [6]. Unsupervised approaches have been proposed as well, however they applied to shallow semantic tasks [14]. 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 [7] and other forms of shallow semantic processing, our approach maps text to formal meaning representations, obtained via generalization.

There are a number of applications of formal concepts in building natural language taxonomies. Formal framework based on formal concept lattices that categorizes epistemic communities automatically and hierarchically, rebuilding a relevant taxonomy in the form of a hypergraph of epistemic sub-communities, has been proposed in [23]. The study of concepts can advance further by clarifying the meanings of basic terms such as "prototype" and by constructing a large-scale primary taxonomy of concept types [11]. Based on concept structures, two secondary concept taxonomies and one of conceptual structures has been built, where the primary taxonomy 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. [1] 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.

The use of syntactic generalization in this work is two-fold. Firstly, it is used off-line to form the node of taxonomy tree, finding commonalities between search results for a given taxonomy 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 taxonomy-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 a vertical domain.

**Acknowledgements.** This research is funded by the European Union project Num. 238887, *a unique European citizens' attention service* (iSAC6+) IST-PSP, the ACC1Ó Catalan Government *grant ASKS - Agents for Social Knowledge Search*, the Spanish MCINN (Ministerio de Ciencia e Innovación) project IPT-430000-2010-13 project *Social powered Agents for Knowledge search Engine (SAKE),* and the CSI-ref.2009SGR-1202.

## References

1. Alani, H., Brewster, C.: Ontology ranking based on the analysis of concept structures. In: K-CAP 2005 Proceedings of the 3rd International Conference on Knowledge Capture (2005)
2. Heddon, H.: Better Living Through Taxonomies. Digital Web Magazine (2008), `http://www.digital-web.com/articles/better_living_through_taxonomies/`
3. Allen, J.F.: Natural Language Understanding, Benjamin Cummings (1987)
4. Chakrabarti, D., Faloutsos, C.: Graph Mining: Laws, Generators, and Algorithms. ACM Computing Surveys 38(1) (2006)
5. Dzikovska, M., Swift, M., Allen, J., de Beaumont, W.: Generic parsing for multi-domain semantic interpretation. In: International Workshop on Parsing Technologies (IWPT 2005), Vancouver BC (2005)
6. Cardie, C., Mooney, R.J.: Machine Learning and Natural Language. Machine Learning 1(5) (1999)
7. Carreras, X., Marquez, L.: Introduction to the CoNLL-2004 shared task: Semantic role labeling. In: Proceedings of the Eighth Conference on Computational Natural Language Learning, pp. 89–97. ACL, Boston (2004)
8. Galitsky, B.: Natural Language Question Answering System: Technique of Semantic Headers. In: Advanced Knowledge International, Australia (2003)
9. Galitsky, B., Dobrocsi, G., de la Rosa, J.L., Kuznetsov, S.O.: From Generalization of Syntactic Parse Trees to Conceptual Graphs. In: Croitoru, M., Ferré, S., Lukose, D. (eds.) ICCS 2010. LNCS, vol. 6208, pp. 185–190. Springer, Heidelberg (2010)

10. Galitsky, B.: Disambiguation Via Default Rules Under Answering Complex Questions. Intl. J. AI. Tools 14(1-2) (2005)
11. Howard, R.W.: Classifying types of concept and conceptual structure: Some taxonomies. Journal of Cognitive Psychology 4(2), 81–111 (1992)
12. Plotkin., G.D.: A note on inductive generalization. In: Meltzer, Michie (eds.) Machine Intelligence, vol. 5, pp. 153–163. Edinburgh University Press, Edinburgh (1970)
13. Ravichandran, D., Hovy, E.: Learning surface text patterns for a Question Answering system. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002), Philadelphia, PA (2002)
14. Lin, D., Pantel, P.: DIRT: discovery of inference rules from text. In: Proc. of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2001, pp. 323–328 (2001)
15. Durme, B.V., Huang, Y., Kupsc, A., Nyberg, E.: Towards light semantic processing for question answering. In: HLT Workshop on Text Meaning (2003)
16. Kapoor, S., Ramesh, H.: Algorithms for Enumerating All Spanning Trees of Undirected and Weighted Graphs. SIAM J. Computing 24, 247–265 (1995)
17. De la Rosa, J.L., Rovira, M., Beer, M., Montaner, M., Gibovic, D.: Reducing Administrative Burden by Online Information and Referral Services. In: Reddick, C.G. (ed.) Citizens and E-Government: Evaluating Policy and Management, pp. 131–157. IGI Global, Austin (2010)
18. López Arjona, A.M., Rigall, M.M., de la Rosa i Esteva, J.L., Regàs, M.M.R.I.: POP2.0: A search engine for public information services in local government. In: Angulo, C., Godo, L. (eds.) Artificial Intelligence Research and Development, vol. 163, pp. 255–262. IOS Press, Amsterdam (2007)
19. Kozareva, Z., Hovy, E., Riloff, E.: Learning and Evaluating the Content and Structure of a Term Taxonomy. In: Learning by Reading and Learning to Read AAAI Spring Symposium, Stanford CA (2009)
20. Liu, J., Birnbaum, L.: What do they think? Aggregating local views about news events and topics. In: WWW 2008, pp. 1021–1022 (2008)
21. Liu, J., Birnbaum, L.: Measuring Semantic Similarity between Named Entities by Searching the Web Directory. Web Intelligence, 461–465 (2007)
22. Kerschberg, L., Kim, W., Scime, A.: A Semantic Taxonomy-Based Personalizable Meta-Search Agent. In: Truszkowski, W., Hinchey, M., Rouff, C.A. (eds.) WRAC 2002. LNCS, vol. 2564, pp. 3–31. Springer, Heidelberg (2003)
23. Roth, C.: Compact, evolving community taxonomies using concept lattices ICCS 14, July 17-21, Aalborg, DK (2006)