Incorporating Trustiness and Collective Synonym/Contrastive Evidence into Taxonomy Construction

Luu Anh Tuan #1, **Jung-jae Kim** #2, **Ng See Kiong** *3

#*School of Computer Engineering, Nanyang Technological University, Singapore*
1anhtuan001@e.ntu.edu.sg, 2jungjae.kim@ntu.edu.sg
**Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore*
3skng@i2r.a-star.edu.sg

Abstract

Taxonomy plays an important role in many applications by organizing domain knowledge into a hierarchy of *is-a* relations between terms. Previous works on the taxonomic relation identification from text corpora lack in two aspects: 1) They do not consider the trustiness of individual source texts, which is important to filter out incorrect relations from unreliable sources. 2) They also do not consider collective evidence from synonyms and contrastive terms, where synonyms may provide additional supports to taxonomic relations, while contrastive terms may contradict them. In this paper, we present a method of taxonomic relation identification that incorporates the trustiness of source texts measured with such techniques as PageRank and knowledge-based trust, and the collective evidence of synonyms and contrastive terms identified by linguistic pattern matching and machine learning. The experimental results show that the proposed features can consistently improve performance up to 4%-10% of F-measure.

1 Introduction

Taxonomies which serve as backbone of structured knowledge are useful for many applications such as question answering (Harabagiu et al., 2003) and document clustering (Fodeh et al., 2011). Even though there are many hand-crafted, well-structured taxonomies publicly available, including WordNet (Miller, 1995), OpenCyc (Matuszek et al., 2006), and Freebase (Bollacker et al., 2008), they are incomplete in specific domains, and it is time-consuming to manually extend them or create new ones. There is thus a need for automatically extracting taxonomic relations from text corpora to construct/extend taxonomies.

Previous works on the task of taxonomy construction capture information about potential taxonomic relations between concepts, rank the candidate relations based on the captured information, and integrate the highly ranked relations into a taxonomic structure. They utilize such information as hypernym patterns (e.g. A is a B, A such as B) (Kozareva et al., 2008), syntactic dependency (Drumond and Girardi, 2010), definition sentences (Navigli et al., 2011), co-occurrence (Zhu et al., 2013), syntactic contextual similarity (Tuan et al., 2014), and sibling relations (Bansal et al., 2014).

They, however, lack in the three following aspects: 1) Trustiness: Not all sources are trustworthy (e.g. gossip, forum posts written by non-experts) (Dong et al., 2015). The trustiness of source texts is important in taxonomic relation identification because evidence from unreliable sources can be incorrect. For example, the invalid taxonomic relation between "American chameleon" and "chameleon" is mistakenly more popular in the Web than the valid taxonomic relation between "American chameleon" and "lizard", and statistical methods without considering the trustiness may incorrectly extract the invalid relation instead of the latter. However, to the best of our knowledge, no previous work considered this aspect.

2) Synonyms: A concept may be expressed in multiple ways, for example with synonyms. The previous works mostly assumed that a term represents an independent concept, and did not combine information about a concept, which is expressed with multiple synonyms. The lack of evidence from synonyms may hamper the ranking of candidate taxonomic relations. Navigli and Velardi (2004) combined synonyms into a concept, but only for those from WordNet, called synsets.

3) Contrastive terms: We observe that if two terms are often contrasted (e.g. A but not B, A is different from B) (Kim et al., 2006), they may

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1025–1034,
Lisbon, Portugal, 17-21 September 2015. ©2015 Association for Computational Linguistics.

not have a taxonomic relation.

In this paper, we present a method based on the state-of-the-art method (Tuan et al., 2014), which incorporates the trustiness of source texts and the collective evidence from synonyms/contrastive terms. Tuan et al. (2014) rank candidate taxonomic relations based on such evidence as hypernym patterns, WordNet, and syntactic contextual similarity, where the pattern matches and the syntactic contexts are found from the Web by using a Web search engine. First, we calculate the trustiness score of each data source with the four following weights: importance (if it is linked by many pages), popularity (if it is visited by many users), authority (if it is from a creditable Web site) and accuracy (if it has many facts). We integrate this score to the work of (Tuan et al., 2014) so that evidence for taxonomic relations from unreliable sources is discarded.

Second, we identify synonyms of two terms (t_1, t_2), whose taxonomic relation is being scrutinized, by matching such queries as "t_1 also known as" against the Web to find t_1's synonyms next to the query matches (e.g. t in "t_1 also known as t"). We then collect the evidence for all taxonomic relations between t'_1 and t'_2, where t'_i is either t_i or its synonym ($i \in \{1, 2\}$), and combine them to calculate the evidence score of the candidate taxonomic relation between t_1 and t_2. Similarly, for each pair of two terms (t_1, t_2), we collect their contrastive evidence by matching such queries as "t_1 is not a type of t_2" against the Web, and use them to proportionally decrease the evidence score for taxonomic relation between contrasting terms.

2 Related Work

The previous methods for identifying taxonomic relations from text can be generally classified into two categories: linguistic and statistical approaches. The former approach mostly exploits lexical-syntactic patterns (e.g. *A is a B*, *A such as B*) (Hearst, 1992). Those patterns can be manually created (Kozareva et al., 2008; Wentao et al., 2012) or automatically identified (Navigli et al., 2011; Bansal et al., 2014). The pattern matching methods show high precision when the patterns are carefully defined, but low coverage due to the lack of contextual analysis across sentences.

The latter approach, on the other hand, includes asymmetrical term co-occurrence (Fotzo and Gallinari, 2004), clustering (Wong et al., 2007), syntactic contextual similarity (Tuan et al., 2014), and word embedding (Fu et al., 2014). The main idea behind these techniques is that the terms that are asymmetrically similar to each other with regard to, for example, co-occurrences, syntactic contexts, and latent vector representation may have taxonomic relationships. Such methods, however, usually suffer from low accuracy, though showing relatively high coverage. To get the balance between the two approaches, Yang and Callan (2009), Zhu et al. (2013) and Tuan et al. (2014) combine both statistical and linguistic features in the process of finding taxonomic relations.

Most of these previous methods do not consider if the source text of evidence (e.g. co-occurrences, pattern matches) is trustworthy or not and do not combine evidence from synonyms and contrastive terms as discussed earlier. Related to synonyms, a few previous works utilize siblings for taxonomy construction. Yang and Callan (2009) use siblings as one of the features in the metric-based framework which incrementally clusters terms to form taxonomies. Wentao et al. (2012) also utilize such sibling feature that, for example of the linguistic pattern "A such as B_1, B_2, $\cdots$ and B_n", if the concept at the k-th position (e.g. B_k) from pattern keywords (e.g. such as) is a valid sub-concept (e.g. of A), then most likely its siblings from position 1 to position k-1 (e.g. B_1, $\cdots$, B_{k-1}) are also valid sub-concepts. Bansal et al. (2014) include the sibling factors to a structured probabilistic model over the full space of taxonomy trees, thus helping to add more evidence to taxonomic relations. Navigli and Velardi (2004) utilize the synonym feature (i.e. WordNet synsets) for the process of semantic disambiguation and concept clustering as mentioned above, but not for the process of inducing novel taxonomic relations.

3 Methodology

We briefly introduce (Tuan et al., 2014) in Section 3.1. We then explain how to incorporate trustiness (Section 3.2) and collective evidence from synonyms (Section 3.3) and from contrastive terms (Section 3.4) into the work of (Tuan et al., 2014).

3.1 Overview of baseline (Tuan et al., 2014)

Tuan et al. (2014) follow three steps to construct a taxonomy: term extraction/filtering, taxonomic relation identification and taxonomy induction. Because the focus of this paper is on the second step, taxonomic relation identification, we use the same methods for the first and third steps as in (Tuan et al., 2014) and will not discuss them here.

Given each ordered pair of two terms t_1 and t_2 from the term extraction/filtering, the taxonomic relation identification of (Tuan et al., 2014) calculates the evidence score that t_1 is a hypernym of t_2 (denoted as $t_1 \gg t_2$) based on the following three measures:

String inclusion with WordNet (SIWN): This measure is to check if t_1 is a substring of t_2, considering synonymy between words using WordNet synsets. $Score_{SIWN}(t_1, t_2)$ is set to 1 if there is such evidence; otherwise, it is set to 0.

Lexical-syntactic pattern (LSP): This measure is to find how much more evidence for $t_1 \gg t_2$ is found in the Web than for $t_2 \gg t_1$. Specifically, a list of manually constructed hypernym patterns (e.g. "t_2 is a t_1") is queried with a Web search engine to estimate the number of evidence for $t_1 \gg t_2$ from the Web. The LSP measure is calculated as follows, where $C_{Web}(t_1, t_2)$ denotes the collection of search results:

$$Score_{LSP}(t_1, t_2) = \frac{log(|C_{Web}(t_1,t_2)|)}{1+log(|C_{Web}(t_2,t_1)|)}$$

Syntactic contextual subsumption (SCS): The idea of this measure is to derive the hypernymy evidence for two terms t_1 and t_2 from their syntactic contexts, particularly from the triples of (subject,verb,object). They observe that if the context set of t_1 mostly contains that of t_2 but not vice versa, then t_1 is likely to be a hypernym of t_2. To implement this idea, they first find the most common relation (or verb) r between t_1 and t_2, and use the queries "t_1 r" and "t_2 r" to construct two corpora $Corpus^{\Gamma}_{t_1}$ and $Corpus^{\Gamma}_{t_2}$ for t_1 and t_2, respectively. Then the syntactic context sets are created from these contextual corpora using a non-taxonomic relation identification method. The details of calculating $Score_{SCS}(t_1, t_2)$ can be found in (Tuan et al., 2014).

They linearly combine the three scores as follows:

$$\begin{aligned} Score(t_1, t_2) = {} & \alpha \times Score_{SIWN}(t_1, t_2) \\ & + \beta \times Score_{LSP}(t_1, t_2) + \gamma \times Score_{SCS}(t_1, t_2) \end{aligned} \quad (1)$$

If $Score(t_1, t_2)$ is greater than a threshold value, then t_1 is regarded as a hypernym of t_2. We use the same values of α, β and γ as in (Tuan et al., 2014).

3.2 Trustiness of the evidence data

We introduce our method of estimating the trustiness of a given source text in Section 3.2.1 and explain how to incorporate it into the work of (Tuan et al., 2014) in Section 3.2.2.

3.2.1 Collecting trustiness score of the evidence data

Given a data source (e.g. Web page), we consider four aspects of trustiness as follows:

- Importance: A data source may be important if it is referenced by many other sources.
- Popularity: If a data source is accessed by many people, it is considered popular.
- Authority: If the data is created by a trusted agency, such as government and education institute, it may be more trustful than others from less trusted sources such as forums and social media.
- Accuracy: If the data contains many pieces of accurate information, it seems trustful.

Importance

To measure the importance of a Web page (d) as data source, we use the Google PageRank score[1] ($Score_{PageRank}(d)$) that is calculated based on the number and quality of links to the page. The PageRank scores have the scale from 0 to 9, where the bigger score means more importance than the lower one. Using this score, the importance of a page is calculated as follows:

$$Trust_{Imp}(d) = \frac{1}{10 - Score_{PageRank}(d)} \quad (2)$$

Note that we use the non-linearity for PageRank score rather than just normalizing PageRank to 0-1. The reason is to vary the gaps between the important sites (which usually have the PageRank score value from 7-10) and majority unimportant site (which usually have the PageRank score value less than 5).

Popularity

We use Alexa's Traffic Rank[2] as the measure of popularity, ($Score_{Alexa}(d)$) which is based on the traffic data provided by users in Alexa's global data panel over a rolling 3 month period. The Traffic Ranks are updated daily. A site's rank is

[1] http://searchengineland.com/what-is-google-pagerank-a-guide-for-searchers-webmasters-11068

[2] http://www.alexa.com/

based on a combined measure of unique visitors and page views. Using this rank, the popularity of a data source is calculated as follows:

$$Trust_{Pop}(d) = \frac{1}{log(Score_{Alexa}(d) + 1)} \tag{3}$$

We use log transform in the popularity score instead of, for example, linear scoring because we want to avoid the bias of the much large gap between the Alexa scores of different sites (e.g. one site has Alexa score 1000, but the other may have score 100,000)

Authority

We rank the authority of a data source based on the internet top-level domain (TLD). We observe that the pages with limited and registered TLD (e.g. *.gov*, *.mil*, *.edu*) are often more credible than those with open domain (e.g. *.com*, *.net*). Therefore, the authority score of a data source is calculated as follows:

$$Trust_{Auth}(d) = \begin{cases} 1 & \text{if TLD of } d \text{ is } \textit{.gov}, \textit{.mil} \text{ or } \textit{.edu} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Note that there are some reasons we choose such implementation of Authority in an elementary way. First, we tried finer categorization of various domains, e.g. *.int* has score 1/3, *.com* has score 1/4, etc. However, the experimental results did not show much change of performance. In addition, there is controversy on which open TLD domains are more trustful than the others, e.g. it is difficult to judge whether a *.net* site is more trustful than *.org* or not. Thus, we let all open TLD domains have the same score.

Accuracy

If the data source contains many pieces of accurate information, it is trustful. Inspired by the idea of Dong et al. (2015), we estimate the accuracy of a data source by identifying correct and incorrect information in form of the triples *(Subject, Predicate, Object)* in the source, where *Subject*, *Predicate* and *Object* are normalized with regard to the knowledge base Freebase. The extraction of the triples includes six tasks: named entity recognition, part of speech tagging, dependency parsing, triple extraction, entity linkage (which maps mentions of proper nouns and their co-references to the corresponding entities in Freebase) and relation linkage. We use three information extraction (IE) tools (Angeli et al. (2014), Manning et al. (2014), MITIE[3]) for the first four tasks, and develop a method similar to Hachey et al. (2013) for the last two tasks of entity linkage and relation linkage.

Since the IE tools may produce noisy or unreliable triples, we use a voting scheme for triple extraction as follows: A triple is only considered to be true if it is extracted by at least two extractors. After obtaining all triples in the data source, we use the closed world assumption as follows: Given subject s and predicate p, $O(s,p)$ denotes the set of such objects that a triple (s,p,o) is found in Freebase. Now given a triple (s,p,o) found in the data source, if $o \in O(s,p)$, we conclude that the triple is correct; but if $o \notin O(s,p)$ and $|O(s,p)| > 0$, we conclude that the triple is incorrect. If $|O(s,p)| = 0$, we do not conclude anything about the triple, and the triple is removed from the set of facts found in the data source.

Given a data source d, we define $cf(d)$ as the number of correct facts, and $icf(d)$ as the number of incorrect facts found in d. The accuracy of d is calculated as follows:

$$Trust_{Accu}(d) = \frac{1}{1 + icf(d)^2} - \frac{1}{1 + cf(d)^2} \tag{5}$$

Combining trustiness scores

The final trustiness score of a data source is the linear combination of the four scores as follows:

$$\begin{aligned} Trust(d) = \alpha \times Trust_{Imp}(d) + \beta \times Trust_{Pop}(d) \\ + \gamma \times Trust_{Auth}(d) + \delta \times Trust_{Accu}(d) \end{aligned} \tag{6}$$

To estimate the optimal combination for parameters α, β, γ and δ, we apply linear regression algorithm (Hastie et al., 2009). For parameter learning, we manually list 50 websites as trusted sources (e.g. *stanford.edu*, *bbc.com*, *nasa.gov*), and the top 15 gossip websites listed in a site[4] as untrusted sources. Then we use the scores of their individual pages by the four methods to learn the parameters in Formula 6. The learning results are as follows: α=0.46, β=0.46, γ=2.03, δ=0.61.

3.2.2 Integrating trustiness into taxonomic relation identification methods

Given a data collection C, we define:

$$AvgTrust(C) = \frac{\sum_{d \in C} Trust(d)}{|C|}$$

[3] https://github.com/mit-nlp/MITIE

[4] http://www.ebizmba.com/articles/gossip-websites

as the average trustiness score of all data in C.

We integrate the trustiness score to the three taxonomic relation identification methods described in Section 3.1 as follows:

LSP method:
The LSP evidence score of the taxonomic relation between t_1 and t_2 is recalculated as follows:

$$Score^{Trust}_{LSP}(t_1, t_2) = Score_{LSP}(t_1, t_2) \times \\ (AvgTrust(C_{Web}(t_1, t_2)) + AvgTrust(C_{Web}(t_2, t_1))) \quad (7)$$

The intuition of Formula 7 is that the original LSP evidence score is multiplied with the average trustiness score of all evidence documents for the taxonomic relation from the Web. If the number of Web search results is too large, we use only the first 1,000 results to estimate the average trustiness score.

SCS method:
Similarly, the SCS evidence score is recalculated as follows:

$$Score^{Trust}_{SCS}(t_1, t_2) = Score_{SCS}(t_1, t_2) \times \\ (AvgTrust(Corpus^{\Gamma}_{t_1}) + AvgTrust(Corpus^{\Gamma}_{t_2})) \quad (8)$$

SIWN method:
This method does not use any evidence from the Web, and so its measure does not change as follows:

$$Score^{Trust}_{SIWN}(t_1, t_2) = Score_{SIWN}(t_1, t_2) \quad (9)$$

The three measures of trustiness are also linearly combined as follows:

$$Score^{Trust}(t_1, t_2) = \alpha \times Score^{Trust}_{SIWN}(t_1, t_2) + \\ \beta \times Score^{Trust}_{LSP}(t_1, t_2) + \gamma \times Score^{Trust}_{SCS}(t_1, t_2) \quad (10)$$

The values of α, β, and γ in Formula 10 are identical to those for Formula 1.

3.3 Collective synonym evidence

3.3.1 Synonymy identification

We use the three following methods to collect synonyms: dictionaries, pattern matching, and supervised learning.

Dictionaries: Synonyms can be found in dictionaries like a general-purpose dictionary WordNet and also domain-specific ones. Since our domains of interest include virus, animals, and plants (see the next section for details), we also utilize MeSH[5], a well-known vocabulary in biomedicine.

Pattern matching: Given two terms t_1 and t_2, we use the following patterns to find their synonymy evidence from the Web:

- t_1 also [known|called|named|abbreviated] as t_2
- Other common name[s] of t_1 [is|are|include] t_2
- t_1, or t_2, is a
- t_1 (short for t_2)

, where $[a|b]$ denotes a choice between a and b. If the number of Web search results is greater than a threshold Ψ, t_1 is considered as a synonym of t_2.

Supervised learning: We randomly pick 100 pairs of synonyms in WordNet, and for each pair, we use the Web search engine to collect sample sentences in which both terms of the pair are mentioned. If the number of collected sentences is greater than 2000, we use only the first 2000 sentences for training. After that, we extract the following features from the sentences to train a logistic regression model (Hastie et al., 2009) for the synonymy identification:

- Headwords of the two terms
- Average distance between the terms
- Sequence of words between the terms
- Bag of words between the terms
- Dependency path between the terms (using Stanford parser (Klein and Manning, 2003))
- Bag of words on the dependency path

The average F-measure of the obtained model with 10-fold cross-validation is 81%. We use the learned model to identify more synonym pairs in the next step.

3.3.2 Embedding synonym information

Given a term t, we denote $Syn(t)$ as the set of synonyms of t (including t itself). The evidence scores of the SCS and LSP methods are recalculated with synonyms as follows:

$$Score^{Synonym}_{X}(t_1, t_2) = \sum_{\substack{t'_1 \in Syn(t_1) \\ t'_2 \in Syn(t_2)}} Score_X(t'_1, t'_2) \quad (11)$$

, where the variable X can be replaced with SCS and LSP.

[5] http://www.ncbi.nlm.nih.gov/mesh

The intuition of Formula (11) is that the evidence score of the taxonomic relation between two terms t_1 and t_2 can be boosted by adding all the evidence scores of taxonomic relations between them and their synonyms.

Again, as for the SIWN method, we do not change the evidence score as follows:

$$Score^{Synonym}_{SIWN}(t_1, t_2) = Score_{SIWN}(t_1, t_2)$$

3.4 Contrastive evidence

Given two terms t_1 and t_2, we use the following patterns to find their contrastive (thus negative) evidence from the Web:

- t_1 is not a t_2
- t_1, unlike the t_2
- t_1 but not t_2
- t_1 is not a [type|kind] of t_2
- t_1 is different [from|with] t_2
- t_1, not t_2

$WH(t_1, t_2)$ denotes the total number of Web search results, and the contrastive evidence score between t_1 and t_2 is computed as follows:

$$Contrast(t_1, t_2) = \log(WH(t_1, t_2) + 1) \tag{12}$$

Similar to the collective synonym evidence, the contrastive evidence score of taxonomic relation between t_1 and t_2 is boosted with the contrastive evidence scores of taxonomic relations between the two terms and their synonyms as follows:

$$Score^{Contrast}(t_1, t_2) = \frac{\sum_{\substack{t'_1 \in Syn(t_1)\\ t'_2 \in Syn(t_2)}} Contrast(t'_1, t'_2)}{|Syn(t_1)| * |Syn(t_2)|} \tag{13}$$

3.5 Combining trustiness, synonym and contrastive evidence

We combine all the three features into the system of (Tuan et al., 2014) as follows:

$$\begin{aligned} Score^{Final}_{X}(t_1, t_2) =& Score^{Trust}_{X}(t_1, t_2) \\ &+ Score^{Synonym}_{X}(t_1, t_2) \end{aligned} \tag{14}$$

, where the variable X can be replaced with each of the three taxonomic relation evidence measures (i.e. SCS, LSP, SIWN). The final combined score is calculated as follows:

$$\begin{aligned} &Score^{Final}_{Combined}(t_1, t_2) = \ \alpha \times Score^{Final}_{SIWN}(t_1, t_2) \\ &\quad + \beta \times Score^{Final}_{LSP}(t_1, t_2) + \ \gamma \times Score^{Final}_{SCS}(t_1, t_2) \\ &\quad - \delta \times Score^{Contrast}(t_1, t_2) \end{aligned} \tag{15}$$

For each ordered pair of terms t_1 and t_2, if $Score^{Final}_{Combined}(t_1, t_2)$ is greater than a threshold value, then t_1 is considered as a hypernym of t_2.

We estimate the optimal values of parameters α, β, γ and δ in Formula 15 with ridge regression technique (Hastie et al., 2009) as follows: First, we randomly select 100 taxonomic relations in Animal domain as the training set. For each taxonomic relation $t_1 \gg t_2$, its evidence score is estimated as $\tau + \frac{1}{Dist(t_1,t_2)}$, where τ is the threshold value for $Score^{Final}_{Combined}$, and $Dist(t_1, t_2)$ is the length of the shortest path between t_1 and t_2 found in WordNet. Then we use our system to find evidence scores with taxonomic relation identification methods in Formulas 13 and 14. Finally, we build the training set using Formula 15, and use the ridge regression algorithm to learn that the best value for α is 1.31, β 1.57, γ 1.24 and δ 0.79, where τ=2.3.

4 Experiment

4.1 Datasets

We evaluate our method for taxonomy construction against the following collections of six domains:

- Artificial Intelligence (AI) domain: The corpus consists of 4,976 papers extracted from the IJCAI proceedings from 1969 to 2014 and the ACL archives from year 1979 to 2014.
- Finance domain: The corpus consists of 1,253 papers extracted from the freely available collection of "Journal of Financial Economics" from 1995 to 2012 and from "Review Of Finance" from 1997 to 2012.
- Virus domain: We submit the query "virus" to PUBMED search engine [6] and retrieve the first 20,000 abstracts as the corpus of the virus domain.
- Animals, Plants and Vehicles domains: Collections of Web pages crawled by using the bootstrapping algorithm described by Kozareva et al. (2008).

We report the results of two experiments in this section: (1) Evaluating the construction of new taxonomies for Finance and AI domains (Section 4.2), and (2) comparing with the curated

[6] http://www.ncbi.nlm.nih.gov/pubmed

databases' sub-hierarchies. We compare our approach with other three state-of-the-art methods in the literature, i.e. (Kozareva and Hovy, 2010), (Navigli et al., 2011) and (Tuan et al., 2014) (Section 4.3). In addition, for Animal domain, we also compare with the reported performance of Bansal et al. (2014), a recent work to construct taxonomy using belief propagation.

4.2 Constructing new taxonomies for Finance and AI domains

Referential taxonomy structures such as WordNet and OpenCyc are widely used in semantic analytics applications. However, their coverage is limited to common, well-known areas, and many specific domains like Finance and AI are not well covered in those structures. Therefore, an automatic method which can induce taxonomies for those specific domains can greatly contribute to the process of knowledge discovery.

To estimate the precision of a given method, we randomly choose 100 relations among the results of the method and manually check their correctness. The results summarized in Table 1 show that our method extracts much more relations, though with slightly lower precision, than Kozareva et al. (2008) and Navigli and Velardi (2004). Note that due to the lack of gold standards in these two domains, we do not compare the methods in terms of F-score, which we will measure with curated databases in the next section. Compared to Tuan et al. (2014), which can be considered as the baseline of our approach, our method has significant improvement in both precision and the number of extracted relations. It indicates that the three incorporated features of trustiness, and synonym and contrastive evidence are effective in improving the performance of existing taxonomy construction methods.

	Finance		AI	
	P	N	P	N
Kozareva	**90%**	753	**94%**	950
Navigli	88%	1161	93%	1711
Tuan	85%	1312	90%	1927
Our method	88%	**1570**	92%	**2168**

Table 1: Experiment result for finance and AI domains. P stands for Precision, and N indicates the number of extracted relations.

4.3 Evaluation against curated databases

We evaluate automatically constructed taxonomies for four domains (i.e. Animal, Plant, Vehicle, Virus) against the corresponding sub-hierarchies of curated databases. For Animal, Plant and Vehicle domains, we use WordNet as the gold standards, whereas for Virus domain, we use MeSH sub-hierarchy of virus as the reference.

Note that in this comparison, to be fair, we change our algorithm to avoid using WordNet in identifying taxonomic relations (i.e. SIWN algorithm), and we only use the exact string-matching comparison without WordNet. The evaluation uses the following measures:

$$Precision = \frac{\#relations\ found\ in\ database\ and\ by\ the\ method}{\#relations\ found\ by\ the\ method}$$

$$Recall = \frac{\#relations\ found\ in\ database\ and\ by\ the\ method}{\#relations\ found\ in\ database}$$

To understand the individual contribution of the three introduced features (i.e. trustiness, synonym, contrast), we also evaluate our method only with one of the three features each time, as well as with all the three features (denoted as "Combined").

Tables 2 and 3 summarize the experiment results. Our combined method achieves significantly better performance than the previous state-of-the-art methods in terms of F-measure and Recall (t-test, p-value $<$ 0.05) for all the four domains. For Animal domain, it also shows higher performance than the reported performance of Bansal et al. (2014). In addition, the proposed method improves the baseline (i.e. Tuan et al. (2014)) up to 4%-10% of F-measure.

Furthermore, we find that the three features have different contribution to the performance improvement. The trustiness feature contributes to the improvement on both precision and recall. The synonym feature has the tendency of improving the recall further than the trustiness, whereas the contrastive evidence improves the precision. Note that we discussed these different contributions of the features in the Introduction.

	Animal			Plant		
	P	R	F	P	R	F
Kozareva	**98**	38	55	**97**	39	56
Navigli	97	44	61	**97**	38	55
Bansal	84	55	67			
Tuan	95	56	70	95	53	68
Trustiness	97	61	75	**97**	56	71
Synonym	92	**65**	76	93	58	71
Contrast	97	55	70	**97**	53	69
Combined	97	**65**	**78**	96	**59**	**73**

Table 2: Experiment results for animal and plant domains. P stands for Precision, R Recall, and F F-score. The unit is %.

	Vehicle			Virus		
	P	R	F	P	R	F
Kozareva	**99**	60	75	97	31	47
Navigli	91	49	64	**99**	37	54
Tuan	93	69	79	93	43	59
Trustiness	96	72	82	97	48	64
Synonym	91	72	80	91	53	67
Contrastive	97	68	80	98	42	59
Combined	95	**73**	**83**	96	**54**	**69**

Table 3: Experiment results for vehicle and virus domains. P stands for Precision, R Recall, and F F-score. The unit is %.

4.3.1 Evaluation of individual methods for trustiness and synonymy identification

We evaluate the individual methods for trustiness measurement and synonymy identification described in Sections 3.2.1 and 3.3.1. For this purpose, we evaluate our system only with one of the individual methods at a time (i.e. importance, popularity, authority and accuracy for trustiness measure, and dictionary, pattern matching, and machine learning methods for synonymy identification).

As summarized in Table 4, the "Importance" and "Accuracy" methods for trustiness measurement based on PageRank and IE systems, respectively, have more contribution than the others. Similarly, the experiment results indicate that the "Machine learning" method has the most contribution among the three methods of synonymy identification.

In addition, we also examine the interdependence of the four introduced aspects of trustiness by running the system with the combination of only two aspects, Importance and Accuracy. The results in all domains show that when combining only the Importance and Accuracy, the system almost achieves the same performance to that of the combined system with all four criteria, except for the Plant domain. It can be explained as the Importance aspect (which is expressed as the PageRank score) may subsume the Popularity and Authority aspects. Another interesting point is that the performance of Accuracy, which is solely based on the local information from the website, when applied individually, is almost the same with that of Importance which is based on the distributed information. It shows that the method of ranking of the sites based on the knowledge-based facts can achieve the effectiveness as good as the traditional ranking method using PageRank score.

	Animal	Plant	Vehicle	Virus
Trustiness:				
Importance	74%	**70%**	81%	63%
Popularity	72%	69%	81%	61%
Authority	72%	69%	80%	61%
Accuracy	73%	**70%**	81%	62%
Imp + Accu	**75%**	**70%**	**82%**	**64%**
Synonym:				
Dictionaries	73%	69%	79%	62%
Pattern matching	**74%**	69%	**80%**	64%
Machine learning	**74%**	**70%**	**80%**	**65%**

Table 4: Contribution of individual trustiness measures and collective synonym evidence in terms of F-measure. Imp stands for Important and Accu stands for Accuracy

4.4 Discussion

4.4.1 Case studies

We give two examples to illustrate how the proposed features help to infer correct taxonomic relations and filter out incorrect relations. Our baseline (Tuan et al. (2014)) extracts an incorrect taxonomic relation between 'fox' and 'flying fox' due to the following reasons: (1) 'flying fox' includes 'fox' (SIWN) and (2) untrusted sources such as a public forum[7] support the relation. Using our proposed method, this relation is filtered out because those untrusted sources are discouraged by the trustiness feature, and also because there are contrastive evidence[8] saying that 'flying fox' is NOT a 'fox'. Specifically, the average trustiness score of LSP method of the sources for the invalid relation (i.e. $AvgTrust(C_{Web}(fox, flying\ fox)) + AvgTrust(C_{Web}(flying\ fox, fox))$) is 0.63, which is lower than the average of those scores, 0.90. Also, the collective contrastive evidence score (i.e. $Score^{Contrast}(fox, flying\ fox)$) is 1.10, which is higher than the average collective contrastive score, 0.32.

On the other hand, the true taxonomic relation between 'bat' and 'flying fox' is not identified by the baseline, mainly due to the rare mention of this relation in the Web. However, our proposed method can recognize this relation because of two reasons: (1) 'flying fox' has many synonyms such as 'fruit bat', 'pteropus', 'kalong', and 'megabat', and there are much evidence that these synonyms are kinds of 'bat' (i.e. using the collective synonym evidence). (2) The evidence for the taxonomic relation between 'fly-

[7] http://redwall.wikia.com/wiki/User:Ferretmaiden/Archive3

[8] http://en.cc-english.com/index.php?shownews-1397

ing fox' and 'bat', though rare, is from trusted sites[9] which are maintained by scientists. Thus, the trustiness feature helps to boost the evidence score for this relation over the threshold value. Specifically, the average trustiness score of LSP method (i.e. $(AvgTrust(C_{Web}(bat, flying\ fox)) + AvgTrust(C_{Web}(flying\ fox, bat))))$, 2.84, is higher than the average in total, 0.90.

We further investigate on 256 taxonomic relations that were missed by the baseline but correctly identified by the proposed method. The average $Score_{LSP}$ and the average $Score_{SCS}$ of the relations by the baseline are 0.35 and 0.60, respectively, while those by the proposed method are 1.17 and 0.82, respectively. We thus find that the proposed method is more effective in correctly improving the LSP method than the SCS method.

4.4.2 Empirical comparison with WordNet

By error analysis, we find that our results may complement WordNet. For example, in Animal domain, our method identifies 'wild sheep' as a hyponym of 'sheep', but in WordNet, they are siblings. However, many references [10] [11] consider 'wild sheep' as a species of 'sheep'. Another such example is that our system recognizes 'aquatic vertebrate' as a hypernym of 'aquatic mammal', but WordNet places them in different subtrees incorrectly [12]. Therefore, our results may help restructure and extend WordNet.

4.4.3 Threshold tuning

Our scoring methods utilize several thresholds to select relations of high ranks. Here we discuss them in details below.

The threshold value Ψ for the pattern matching method in Section 3.3.1 controls the number of synonymy relations extracted from text. The threshold value for $Score_{Combined}^{Final}$ of Formula 15 in Section 3.5 controls the number of extracted taxonomic relations. In general, the larger these threshold values are, the higher number of synonyms and taxonomic relations we can get. In our experiments, we found that the threshold values for Ψ between 100 and 120, and those for $Score_{Combined}^{Final}$ between 2.3 and 2.5 generally help the system achieve the best performance.

5 Conclusion

In this paper, we propose the features of trustiness, and synonym and contrastive collective evidence for the task of taxonomy construction, and show that these features help the system improve the performance significantly. As future work, we will investigate into the task of automatically constructing patterns for the pattern matching methods in Sections 3.3 and 3.4, to improve coverage. We will also enhance the accuracy measure of trustiness, based on the observation that some untrusted sites copy information from other sites to make them look more trustful.

References

Gabor Angeli, Julie Tibshirani, Jean Y. Wu, and Christopher D. Manning. 2014. Combining Distant and Partial Supervision for Relation Extraction. *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, pages 1556–1567.

Mohit Bansal, David Burkett, Gerard De Melo, and Dan Klein. 2014. Structured learning for taxonomy induction with belief propagation. *Proceedings of the 52nd Annual Meeting of the ACL*, pages 1041–1051.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 1247–1250.

Xin L. Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. 2015. Knowledge-Based Trust: Estimating the Trustworthiness of Web Sources. *Proceedings of the VLDB Endowment*, 8(9).

Lucas Drumond and Rosario Girardi. 2010. Extracting ontology concept hierarchies from text using markov logic. *Proceedings of the ACM Symposium on Applied Computing*, pages 1354–1358.

Samah Fodeh, Bill Punch, and Pang N. Tan. 2011. On ontology-driven document clustering using core semantic features. *Knowledge and information systems*, 28(2):395–421.

Hermine N. Fotzo and Patrick Gallinari. 2004. Learning "generalization/specialization" relations between concepts - application for automatically building thematic document hierarchies. *Proceedings of the 7th International Conference on Computer-Assisted Information Retrieval.*

[9] http://krjsoutheastasianrainforests.weebly.com/animals-in-biome-and-habitat-structures.html

[10] http://en.wikipedia.org/wiki/Ovis

[11] http://www.bjornefabrikken.no/side/norwegian-sheep/

[12] http://en.wikipedia.org/wiki/Aquatic_mammal

Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning semantic hierarchies via word embeddings. *Proceedings of the 52nd Annual Meeting of the ACL*, pages 1199–1209.

Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James R. Curran. 2013. Evaluating entity linking with Wikipedia. *Artificial intelligence*, 194:130–150.

Sanda M. Harabagiu, Steven J. Maiorano, and Marius A. Pasca. 2003. Open-domain textual question answering techniques. *Natural Language Engineering*, 9(3):231–267.

Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. 2009. *The elements of statistical learning*. Springer-Verlag.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. *Proceedings of the 14th Conference on Computational Linguistics*, pages 539–545.

Jung J. Kim, Zhuo Zhang, Jong C. Park, and See K. Ng. 2006. Biocontrasts: extracting and exploiting protein-protein contrastive relations from biomedical literature. *Bioinformatics*, 22(5):597–605.

Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. *Proceedings of the 41st Annual Meeting of the ACL*, pages 423–430.

Zornitsa Kozareva and Eduard Hovy. 2010. A Semi-supervised Method to Learn and Construct Taxonomies Using the Web. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1110–1118.

Zornitsa Kozareva, Ellen Riloff, and Eduard H. Hovy. 2008. Semantic Class Learning from the Web with Hyponym Pattern Linkage Graphs. *Proceedings of the 46th Annual Meeting of the ACL*, pages 1048–1056.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. *Proceedings of the 52nd Annual Meeting of the ACL*, pages 55–60.

Cynthia Matuszek, John Cabral, Michael J. Witbrock, and John DeOliveira. 2006. An introduction to the syntax and content of cyc. *Proceedings of the AAAI Spring Symposium*, pages 44–49.

George A. Miller. 1995. WordNet: a Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

Roberto Navigli and Paola Velardi. 2004. Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. *Computational Linguistics*, 30(2):151–179.

Roberto Navigli, Paola Velardi, and Stefano Faralli. 2011. A Graph-based Algorithm for Inducing Lexical Taxonomies from Scratch. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1872–1877.

Luu A. Tuan, Jung J. Kim, and See K. Ng. 2014. Taxonomy Construction using Syntactic Contextual Evidence. *Proceedings of the EMNLP conference*, pages 810–819.

Wu Wentao, Li Hongsong, Wang Haixun, and Kenny. Q. Zhu. 2012. Probase: A probabilistic taxonomy for text understanding. *Proceedings of the ACM SIGMOD conference*, pages 481–492.

Wilson Wong, Wei Liu, and Mohammed Bennamoun. 2007. Tree-traversing ant algorithm for term clustering based on featureless similarities. *Data Mining and Knowledge Discovery*, 15(3):349–381.

Hui Yang and Jamie Callan. 2009. A Metric-based Framework for Automatic Taxonomy Induction. *Proceedings of the 47th Annual Meeting of the ACL*, pages 271–279.

Xingwei Zhu, Zhao Y. Ming, and Tat S. Chua. 2013. Topic hierarchy construction for the organization of multi-source user generated contents. *Proceedings of the 36th ACM SIGIR conference*, pages 233–242.