Extended Tversky Similarity for Resolving Terminological Heterogeneities across Ontologies

DuyHoa Ngo, Zohra Bellahsene, and Konstantin Todorov

LIRMM, University Montpellier 2
firstname.lastname@lirmm.fr

Abstract. We propose a novel method to compute similarity between cross-ontology concepts based on the amount of overlap of the information content of their labels. We extend Tversky's similarity measure by using the information content of each term within an ontology label both for the similarity computation and for the weight assignment to tokens. The approach is suitable for handling compound labels. Our experiments showed that it outperforms existing terminological similarity measures for the ontology matching task.

1 A Typology of Terminological Heterogeneities

In a reduced view, a terminology can be defined as a collection of symbols, where each symbol evokes a concept and refers to a concrete object in the real world. Often across different ontologies a concept is denoted by identical or highly similar labels. However, in many cases, these labels differ significantly because of different conventions in the naming process – a phenomenon known as terminological heterogeneity. This heterogeneity is understood as any difference in spelling between two given terms or labels which are assumed to refer to the same concept, i.e. that have the same meaning. *Spelling*[1] is about the lexical expression of concepts, the actual string of characters that is used to label them. *Meaning* refers to the definitions of these labels found in a thesaurus or a lexical database. We assume that the meaning defines a concept, thus identical meanings imply identical concepts. With respect to their spelling, two terms can be very similar or totally different and still mean the same thing in a given context. Moreover, there are different degrees to which this (dis)similarity is manifested. Mind that the more complex the labels, the higher the probability of observing a heterogeneity. Therefore, mapping labels that are composed by multiple tokens is harder than mapping one-token labels. We introduce a scale of the orthographical closeness of two terms representing the different heterogeneity types starting from the lightest expression of heterogeneity and ending with the use of entirely distinct labels to denote a given concept.

[1] The term "syntax" is used instead by other authors but we find that slightly abusive to its meaning in linguistics.

R. Meersman et al. (Eds.): OTM 2013, LNCS 8185, pp. 711–718, 2013.

Terminological heterogeneity types

1. (Almost) identical labels
   - example: in the presence of typos, the use of plural versus singular, etc.
2. Token-wise similarity
   - example: "SumbissionDeadline" vs. "Deadline_Submission".
3. Partial token-wise similarity
   - example: a label is a token of another label, e.g., "Document" vs. "ConferenceDocument".
4. Acronyms / Abbreviations
   - example: “WWW” vs. “WorldWideWeb”, “Misc.” vs. “Miscellaneous”. Token-wise: "MiscTopics" vs. "MiscellaneousTopics".
5. Synonyms
   - example: “booklet” vs. “brochure” (general), “Article” vs. “Paper” (domain specific), “ConferenceDinner” vs. “ConventionBanquet” (token-wise).

Although the heterogeneity types are expressed on orthographical level only, in order to find associations between terms one often needs to apply not only purely spelling-based measures (which would help for heterogeneity type 1), but also semantic similarity measures in order to identify that two in appearance different terms have the same meaning (as in heterogeneity type 5). Finally, mind that linguistic heterogeneity (labels in different natural languages) is not included in this typology, since it goes way beyond spelling mismatches.

Efforts have been made to adapt existing terminological similarity measures from other fields to the ontology matching (OM) task, but there still remain heterogeneities that existing approaches are not able to deal with, especially in the presence of compound concept labels. We address this issue by defining and applying a similarity measure based on techniques coming from the field of information retrieval (IR). More precisely, we make use of the information content (IC) of the tokens composing each label w.r.t. a given ontology. We extend Tversky's similarity measure by using an IC-based weighting of the tokens forming a label.

The paper is structured as follows. We proceed to discuss basic and advanced terminology similarity measures, which have been applied for OM, in Section 2. They are related to the method that we propose in Section 3 and that is defended experimentally in Section 4.

2 Terminological Similarity Measures

In what follows, we assume basic knowledge of the reader on terminological similarity measures for OM [2]. We will use little space for introducing the different measures and focus instead on their strengths and weaknesses in relation to the OM task and the typology presented above.

2.1 Basic Similarity Measures

Here, we consider string-based and language-based measures [2]. *String-based measures* make use of information only relevant to spelling. Following [1], they can be split into edit-based and token-based measures. *Edit-based similarity* of two strings is based on the count of the edit operations required to transform one string into the other (e.g., **Levenstein**). A recent extension called **ISUB** [6] considers not only the similar but also the different parts of two strings. *Token-based similarity* measures compute similarity by splitting the strings into tokens, comparing the tokens by the help of an internal measure and giving the overall similarity of two strings seen as two collections of tokens (e.g., **Q-Grams**). *Language-based measures* rely on linguistic information in order to find the association between terms. They can rely only on the internal linguistic properties of words, or make use of external knowledge resources such as dictionaries, lexicons or thesauri, and look into the semantic relationships in the respective hierarchies.

Discussion. The main advantage of token-based over edit-based measures is their capability to handle compound labels, since they are less sensitive to word swaps (e.g., "MemberConference" and "ConferenceMember"). To overcome tiny variations (e.g., typos), a token-based measure uses an edit-based one as an internal measure. Therefore, token-based measures can be used to deal with heterogeneity type 2, while edit-based measures are appropriate for handling type 1 heterogeneities. Several token-based measures like **TFIDF**, **Jensen-Shannon** and **Fellegi-Sunter** [1] need an external resource to assign weights to tokens. They face the limitation of relying on a large corpus related to a given domain that may not always be available, especially in the OM field. None of these measures can deal with heterogeneity types 3, 4 or 5, since they compute similarity by using spelling-related features only. To overcome heterogeneity type 5, hybrid measures have been proposed (a combination of string-based and language-based similarities). Heterogeneity types 3 and 4 are more difficult to handle and require the use of external knowledge.

Language-based measures, both intrinsic and extrinsic, suffer from two common problems. First, they mainly deal with single words and not with compound labels. For example, neither of the labels "DoctoralThesis" or "PhdDissertation" is found in WordNet although each of their tokens is. The second problem appears when the input words cannot be found in the dictionary due to typos. Therefore, although language-based measures are appropriate to deal with heterogeneities of type 5, they need to be supported by string-based measures.

2.2 Advanced Similarity Measures

Hybrid similarity measures are a combination of string-based and language-based similarity measures. In particular, a hybrid measure can be applied on two levels: between tokens and between compound labels.

One-token labels are aligned by using morphological methods that look at all possible basic forms of each of the two tokens in a dictionary. If the basic forms of both tokens exist, an extrinsic similarity measure is used. Otherwise, the similarity score is computed by a string-based measure. *Compound labels* are first

split into sets of tokens. Having the similarity scores for every pair of tokens, we can apply one of the two widely used aggregation methods, **ExtendedJaccard** and **Monge-Eklan** [3], or the **SoftTFIDF** measure [1].

Discussion. Hybrid similarity measures can be used to deal with both type 2 and type 5 of terminological heterogeneity. When two strings have a high number of shared tokens, which are highly similar in spelling or in meaning, the hybrid similarity measure can detect them as a match. But if the number of shared tokens of two strings is small, both token-based and hybrid measures return a low similarity value and they possibly detect these strings as unmatched. Therefore, for type 3 of terminological heterogeneity, we need to exploit other feature information of entities in order to discover mappings between them.

Weighted techniques have several downsides. In SoftTFIDF, a weight is computed by using the TFIDF approach which requires a large corpus – rarely available for a given matching scenario in the OM world. The Extended Jaccard similarity lacks discriminating power. It would assign similar scores to pairs of tokens which have different relations in a semantic network. For example, $ExtendedJaccard(Publication, Magazine) = ExtendedJaccard(Publication, Journal) = ExtendedJaccard(Publication, Periodical) = 1.0$, although according to WordNet, "journal" and "magazine" are siblings and they are both children of "periodical". Finally, asymmetry appears to be a serious drawback of the SoftTFIDF. To overcome these weaknesses, we have designed a (symmetric) similarity measure extending Tversky's method.

3 Extending Tversky's Similarity Measure

In an ontology which represents a given aspect of the knowledge of a specific domain, certain (non-stop) words frequently appear together with other words in concept labels. For example, in the **conference.owl** ontology, which models the conference organization domain, the total number of concepts is 60. The labels of 14 of these concepts contain the word "conference", and 10 contain the word "contribution", whereas, other words like "author" and "speaker" appear only once as a part of a concept label. Therefore, if the words "conference" and "contribution" are found in a compound label, they are unlikely to be keywords. Instead, they are used to emphasize the specific meaning of the associated words in the domain of interest and disambiguate the meaning of the associated words in different domains.

The proposed measure is inspired by the comparison methods of documents in the IR field and will be applied to deal with type 3 of terminological heterogeneity. After stop-words removal, a weighting method is used to assign a weight to each remaining word which represents the relative importance of that word in the document. Finally, a computation method is applied to calculate a similarity score between two documents.

The main difference between label comparison in OM and generic document comparison in IR is that the former is a comparison of short strings, whereas, the latter is a comparison of long texts. In the OM domain there cannot be found a large corpus from which we can extract the necessary statistical information

for the similarity computation. Therefore, the techniques used in comparison of documents have to be adapted to the label comparison task. In particular, we are going to discuss weight assignment and similarity computation issues which are strongly related to one other.

Weight Assignment. There are many weight assignment approaches proposed in the IR literature (TF, IDF, TFIDF), mainly based on the frequency of occurrence of each word in a document and in a large corpus. In OM, in the first place, there is a lack of a large corpus (a large number of ontologies describing the same domain). Commonly, only two ontologies in a matching scenario are given. Moreover, because of their high heterogeneity, ontologies may slightly overlap or may be totally disjoint w.r.t. terminology. Thus, the words used in one ontology may differ from those used in the other. Consequently, there may be no benefit of calculating the frequency of words across multiple ontologies. In the second place, the weight of a word depends on the ontology that contains that word. As mentioned above, common words in a specific domain may explicitly appear many times in one ontology. They also may not appear but be implicitly represented in the other ontologies. Therefore, if we take multiple ontologies into account, the frequency of occurrence of the common words and keywords may not be significantly different. Consequently, there is not much difference between common words and keywords in the way they are handled by the similarity computation approach.

In our method, a normalization of the IC of each word appearing in an ontology is considered as its weight. In information theory [5], the IC of an object is inversely proportional to the probability of occurrence of that object. We give the IC of a word t appearing in a label as $IC(t) = \log \frac{|T|}{|N|}$, where, $|T|$ is a total number of concepts in a given ontology and $|N|$ is the number of concepts whose label contains t. On this basis, a word t is assigned a weight as follows:

$$weight(t) = \frac{IC(t)}{\max_{i=1..|T|}\{IC(t_i)\}}. \tag{1}$$

Similarity Computation. An appropriate similarity computation method has two of the desired properties described in [6]: *(i) intelligent*: it should recognize the amount of informativeness that each token carries in a label and reflect that on the the similarity score between labels, and *(ii) discriminating*: it should rarely assign the same similarity value when it compares a label to several other similar labels. For example, it should distinguish the similarity of "Publication" to "Journal" from that to "Magazine" and to "Periodical". To fulfill these requirements, the similarity measure should take both the weight values of tokens and the similarity values between tokens into account.

The well-known Tversky similarity measure [7] for two objects A and B seen as sets of features and a function f can be given as: $\mathrm{Tv}(A, B) = \frac{2*f(A\cap B)}{f(A)+f(B)}$. In our case, the objects are labels denoting concepts. Let s_1 and s_2 be two labels and let the function $Tokenize$ return the set of tokens composing a label. Further, let $TokenSim$ be a similarity measure for two terms and let $Share(s_1, s_2) =$

$\{t' \in Tokenize(s_1) \mid \exists t'' \in Tokenize(s_2) \wedge TokenSim(t', t'') \geq \theta\}$. By following Tversky's rule, we give the following definition of the similarity of two labels:

$$\mathrm{ET}(s_1, s_2) = \frac{Common_{s_1,s_2} + Common_{s_2,s_1}}{Total_{s_1} + Total_{s_2}}, \tag{2}$$

where, for $i, j \in \{1, 2\}$ and $i \neq j$, we have

$$Common_{s_i,s_j} = \sum_{t' \in Share(s_i,s_j)} weight(t') \cdot \max_{t'' \in Tokenize(s_j)} (TokenSim(t', t'')),$$

$$Total_{s_i} = \sum_{t \in Tokenize(s_i)} weight(t).$$

The weighting function used in the calculation of the similarity measure can be any function known from the literature. In the particular definition of our similarity measure based on the IC of terms, we have applied the IC-based weight given in Eq. (1). In the next section, we provide a comparison of the outcomes of this measure when different weighting functions are applied.

4 Experiments and Evaluation

Evaluating an OM task consists in comparing the discovered alignments to a reference alignment by the help of evaluation measures corresponding to the harmonic means of Precision (Pr), Recall (Re) and the F-measure (Fm) computed on a set of n tests per matching scenario. A test corresponds to a particular choice of two input ontologies (a source and a target) and a scenario – to a particular matching task (see [4] for details).

We have conducted a series of experiments on two datasets containing terminologically heterogeneous ontologies – the well-known conference dataset from the OAEI[2] and the dataset from the I3CON conference[3]. We have compared our method to basic "weightless" similarity measures and to more advanced similarity measures using a weighting function in the similarity computation. Among the weightless similarity measures, we have chosen the ISUB, Levenstein, Q-Grams, and Monge-Elkan, for reasons of their successful application in the OM field. In addition, each of these measures is representative for its group (string-based, token-based and hybrid, respectively). The weighted measures that we have used are the SoftTFIDF and the Extended Jaccard. As a weighting function for these measures and our Extended Tversky (ET) measure, we have used the IC-based weight proposed in (1) and the standard TFIDF weighting. A mapping selection module is introduced to filter at a given threshold the best candidate mappings. Our results are presented as a function of the different choices of this threshold. They are given in the figures in Tables 1 and 2 for the conference dataset and Tables 3 and 4 for the I3CON dataset.

[2] The Ontology Alignment Evaluation Initiative, `http://oaei.ontologymatching.org`.

[3] `http://www.atl.external.lmco.com/projects/ontology/i3con.html`

Table 1. Conference dataset. Our method is compared to weightless methods by using IC-based weighting (left) and TF-IDF weighting (right).

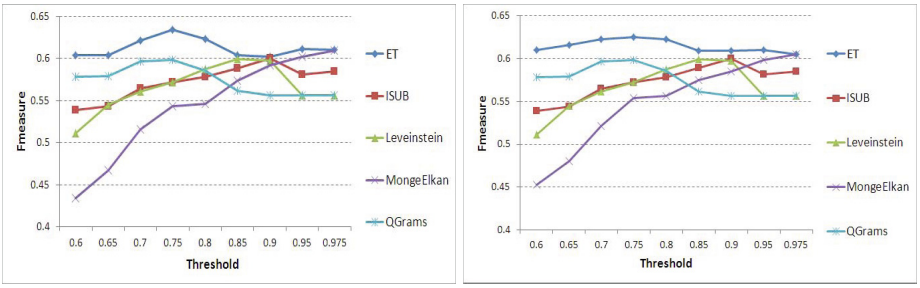
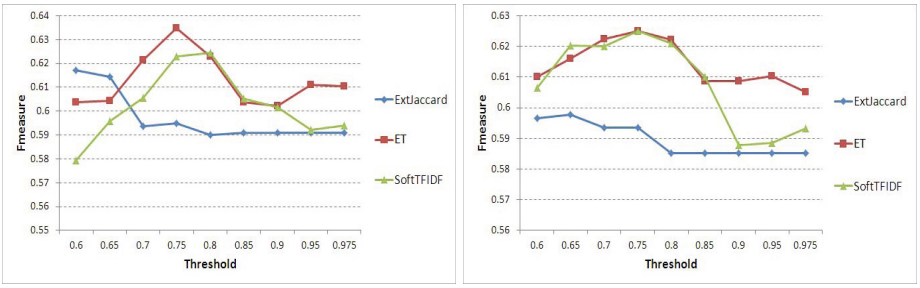


Table 2. Conference dataset. Our method is compared to weighted methods by using IC-based weighting (left) and TF-IDF weighting (right).



As seen in the figures, the ET measure proposed here clearly outperforms the weightless and weighted state-of-the-art measures on the conference dataset for almost all choices of a mapping filter threshold. On the I3CON dataset, we notice that our measure is dominated by the ISUB and the SoftTFIDF measures for certain threshold values. This behavior is explained by the fact that I3CON almost does not contain pairs of heterogeneity type 3. We can see, nevertheless, that the overall highest F-measure values on this dataset are achieved by the ET measure and this measure does not perform globally worse than the other methods on this dataset.

5 Conclusion

The bigger part of the terminological similarity measures that are currently applied to the OM task are borrowed from neighboring domains. However, due to the differences between these domains and the OM domain, these similarity measures need to be adapted in order to perform well. The greatest challenge is to be able to make use in the best possible way of the austere textual information that comes with the ontologies. Addressing this challenge, we have presented a novel similarity measure that is able to deal with certain terminological heterogeneity types in the ontology matching task, that existing techniques cannot handle.

Table 3. I3CON dataset. Our method is compared to weightless methods by using IC-based weighting (left) and TF-IDF weighting (right).

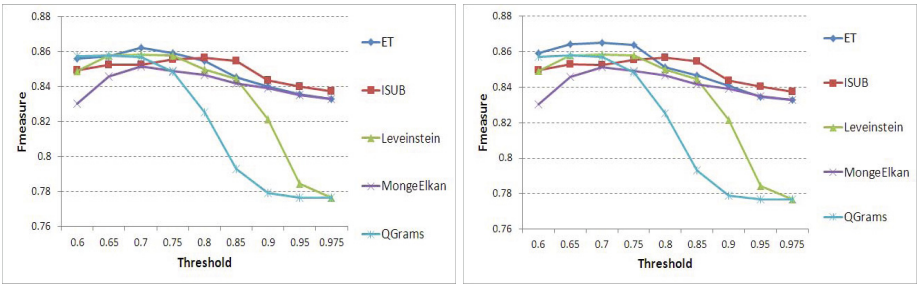
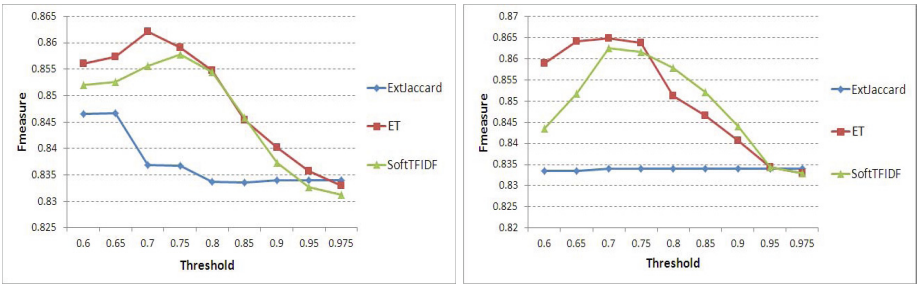


Table 4. I3CON dataset. Our method is compared to weighted methods by using IC-based weighting (left) and TF-IDF weighting (right).



It extends Tversky's similarity and uses the information content of each term within an ontology both for the similarity computation and for the weight assignment to terms. The experimental results show that this similarity measure globally outperforms all existing state-of-the-art techniques, including simple weightless measures and more advanced approaches.

References

1. Cohen, W.W., Ravikumar, P.D., Fienberg, S.E.: A comparison of string distance metrics for name-matching tasks. In: IIWeb, pp. 73–78 (2003)
2. Euzenat, J., Shvaiko, P.: Ontology matching. Springer (2007)
3. Monge, A.E., Elkan, C.P.: An efficient domain-independent algorithm for detecting approximately duplicate database records. In: SIGMOD WS on Research Issues on Data Mining and Knowledge Discovery, pp. 23–29 (1997)
4. Ngo, D., Bellahsene, Z., Todorov, K.: Opening the black box of ontology matching. In: Cimiano, P., Corcho, O., Presutti, V., Hollink, L., Rudolph, S. (eds.) ESWC 2013. LNCS, vol. 7882, pp. 16–30. Springer, Heidelberg (2013)
5. Shannon, C.E.: Prediction and entropy of printed English. Bell Systems Technical Journal, 50–64 (1951)
6. Stoilos, G., Stamou, G., Kollias, S.D.: A string metric for ontology alignment. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 624–637. Springer, Heidelberg (2005)
7. Tversky, A.: Features of similarity. Psychological Review 84, 327–352 (1977)