A Multilingual Test Collection for the Semantic Search of Entity Categories

Juliano Efson Sales[†], Siamak Barzegar[§], Wellington Franco[‡], Bernhard Bermeitinger[†], Tiago Cunha[¶], Brian Davis[§∓], André Freitas[†δ] and Siegfried Handschuh[†]

[†]Department of Computer Science and Mathematics - University of Passau, Germany
[§]Insight Centre for Data Analytics - National University of Ireland, Galway
[‡]Federal University of Ceará, Campus Crateús, Brazil
[¶]University of the International Integration of the Afro-Brazilian Lusophony, Brazil
[∓]Department of Computer Science - Maynooth University, Ireland
[δ]School of Computer Science - The University of Manchester, United Kingdom
{juliano-sales, bernhard.bermeitinger, siegfried.handschuh}@uni-passau.de
siamak.barzegar@insight-centre.org, wellington@crateus.ufc.br
tiagotmc@unilab.edu.br, brian.davis@mu.ie, andre.freitas@manchester.ac.uk

Abstract

Humans naturally organise and classify the world into sets and categories. These categories expressed in natural language are present in all data artefacts from structured to unstructured data and play a fundamental role as tags, dataset predicates or ontology attributes. A better understanding of the category syntactic structure and how to match them semantically is a fundamental problem in the computational linguistics domain. Despite the high popularity of *entity search*, entity categories have not been receiving equivalent attention. This paper aims to present the task of *semantic search of entity categories* by defining, developing and making publicly available a multilingual test collection comprehending English, Portuguese and German. The test collections were designed to meet the demands of the entity search community in providing *more representative* and *semantically complex* query sets. In addition, we also provide comparative baselines and a brief analysis of the results.

Keywords: semantic search, category search, paraphrasing, entity search, multilingual test collection

Lang.	***Category***
EN	American architectural styles
EN	French female artistic gymnasts
EN	Brazilian aerospace conglomerate
PT	Alfabetos derivados do latino
PT	Ordens e congregações religiosas católicas
PT	Primeiros-ministros de Portugal
DE	Ortsteil von Anröchte
DE	Soziale Bewegung als Thema
DE	Teilnehmer an Der Bachelor

Table 1: Examples of categories in English (EN), Portuguese (PT) and German (DE).

1. Introduction

Well-defined tasks and test collections are fundamental resources to allow reproducibility and comparability in information retrieval in general, and entity search in particular (Jones, 1981). As entities are the main target of a vast amount of search queries on the Web (Pound et al., 2010), over the years, a mature research community has emerged around the *entity search* domain.

Despite the considerable number of challenges and campaigns released in the field (Elbedweihy et al., 2015), in recent years, the entity search community has encouraged the development of new tasks. According to Balog and Neumayer (2013) three action points were signed as priorities:

1. (i) getting more representative information needs and favouring long queries over short ones;
2. (ii) limiting search to a smaller, fixed set of entity types (as opposed to arbitrary types of entities); and
3. (iii) using test collections that integrate both structured and unstructured information about entities.

In order to suit at least two of these priorities, this paper aims at presenting the task of *semantic search of entity categories* by defining, developing and making publicly available test collections, and providing a comparative analysis on the baseline results. In addition to English, the test collection comprehends two more languages to contribute to the semantic research community targeting Portuguese and German.

2. The Semantic Search of Entity Categories

Each type or attribute generally describes a singular characteristic of an entity. The nouns *president* and *monument*, the adjective *urban* and named entity *United States* are examples of entity types and attributes found in structured data. Users commonly refer to entities by combining a set of these characteristics to create richer descriptive categories, e.g. *President of the United States* and *Urban monument*. The following text excerpt shows a real example:

> "*Franklin Delano Roosevelt (January 30, 1882 - April 12, 1945), commonly known as FDR, was an American statesman and political leader who served as the 32nd* ***President of the United States****, from 1933 to 1945.*"[1]

President of the United States is a *natural language entity category* labelling the entity *Franklin Delano Roosevelt*.

[1]Extracted from `https://en.wikipedia.org/wiki/Theodore_Roosevelt`

Lang.	*Target entity category*	*Paraphrases*
EN	Irish Manuscripts	Writings from Ireland
EN	Dermatologic Drugs	Pharmaceuticals for the Skin
PT	Países e territórios de língua oficial inglesa	Comunidade anglofônica
PT	Grandes Mestres de xadrez	Enxadristas célebres
DE	Literatur über den Islam	Islamzentrische Bücher
DE	Station der Toronto Subway	Haltestelle der U-Bahn in Toronto

Table 2: Examples of paraphrases in English (EN), Portuguese (PT) and German (DE).

We define entity category as a concise and descriptive structured predicate described in natural language that combines one or more types/attributes of an entity.
In addition to their occurrence in unstructured data, entity categories are also available in the form of structured data. DBpedia (Auer et al., 2007) associates descriptive categories to entities representing in the form of Yago properties (Suchanek et al., 2007), which are available in all languages supported by Wikipedia. This information is extracted from the categories present in the Wikipedia articles which are freely described by the Wikipedia community.
Table 1 presents a list of entity categories extracted from the Yago/DBpedia dataset.
The use of natural language to create such a category allows users to express and potentially search for the same (or close) concept using different words. For example, a ***Brazilian aerospace conglomerate*** is also described as ***Brazilian Planemaker***[2] and categorised as ***Aircraft manufacturers of Brazil***[3]. Having the ability to match paraphrases of complex nominals, allows users to find other relevant categories for different lexical expressions. The current task is designed to evaluate the ability of a system to recognise *paraphrases of entity categories* written in English, Portuguese or German.

2.1. Query Types

According to Pound et al. (2010), in the context of *Entity Search*, there are five types of queries summarised as follows:

- **Entity query**: intending to find a particular entity. Expected results are entities corresponding to some disambiguation of the query entity.
- **Type query**: intending to find entities of a particular type or class. Expected results are entities that are instances of the specified type, or an identifier of the type itself.
- **Attribute query**: intending to find values of a particular attribute of an entity or type. Expected results are the values of an attribute specified in the query.
- **Relation query**: intending to find how two or more entities or types are related. Expected results are the one or more relationships among the query entities or types.
- **Other keyword query**: the query intent is described by some keywords that do not fit into any of the above categories. Expected results are resources providing relevant information.

Complex queries combining characteristics of *type queries* and *attribute queries* can benefit from the use of the natural language categories associated with the entities. The categories can provide a shortcut between the natural language query and the target entities. For example, the INEX-LD Challenge (Wang et al., 2012) defines natural language queries and lists associated entities. For the query "*bicycle sport races*" several relevant entities hold the Yago category *Cycling Competitions*. Intuitively we can assume that *bicycle sport races* and *cycling competitions* are equivalent paraphrases. Creating a mechanism able to pair them makes a shortcut between a natural language expression and a set of entities.
The semantic search of entity categories can be seen as a bridge between unstructured and structured data. In addition, they also suggest which kinds of types/attributes are relevant to create descriptive compositions from the user's point of view. For example, the combination of `dbo:occupation` and `dbo:birthPlace` generally creates *commonly used* descriptive categories (e.g. *French Poets*).

3. Test Collection

The test collections comprehend three knowledge bases of about 345,000 entity categories for English, 105,000 for Portuguese and 235,000 for German, which were created based on the set of DBpedia categories. From these sets, we chose a subset of 110 categories for each language to be part of the query sets. The categories in query sets were chose randomly and filtered later to ensure that they vary in size, number of place/demonym references, number of temporal expressions and different noun phrase components, in order to ensure a high semantic variety in the queries.
Examples of categories in the English query set are:

- categories having different sizes to represent different degrees of word compositionally. For example, *Prehistoric Canines* (two terms); *Victims of Helicopter Accidents or Incidents in the United States* (ten terms).
- categories containing references to places/demonym and temporal expressions. For example, *French Senators Of The Second Empire*; *Political Movements in Italy*.

[2]See the magazine article *Brazilian Planemaker Unveils Its Biggest Military Jet Yet* published by Business Insider.
[3]See the Wikipedia category *Aircraft manufacturers of Brazil*.

Language	*Inter-rater agreement*	*Recall*		*MRR*	
		Top 10	*Top 20*	*Top 10*	*Top 20*
English (EN)	0.9075	0.2806	0.2806	0.3096	0.3143
Portuguese (PT)	0.9489	0.3420	0.3907	0.4641	0.4665
German (DE)	0.9960	0.4487	0.4725	0.6760	0.6768

Table 3: Inter-rater agreement, recall and MRR results considering all relevant categories.

- categories containing different kinds of noun phase components, such as nouns, adjectives, and verbs. For example *Recurring Events Established in 1875*, *Magazines with Year of Establishment Missing*.

We grouped the query sets in blocks of 10 categories and asked volunteers fluent or native in English, Portuguese and German to suggest paraphrases for them. They were instructed to describe the same (or close) meaning using preferably different words and different syntactical structures. After that, we applied a curation process conducted by two researchers to assess the quality of the paraphrases. In the end, we accepted a set of 220 paraphrases for English, 229 for Portuguese and 169 for German, which refer respectively to 98, 96 and 71 categories out of the initial set of 110.

Table 2 shows examples of entity categories and their associated paraphrases proposed by the volunteers in English, Portuguese and German. The task aims at retrieving the original category when querying using its paraphrase.

3.1. Evaluation and Relevance Judgements

The task is defined as an information retrieval problem:

- **Statement**: Let T be a set of entity categories, called target set, and (x, y) a pair of entity categories where $x \in T$, $y \notin T$ and y is a paraphrase of x, meaning that y is a semantic approximation of x, represented by $y \approx x$.
- **Search Procedure**: Let f be a semantic search procedure defined by $L = f(T, y)$ where $L = (T, \geq)$, which means L is an ordered set of entity categories.
- **Evaluation Procedure**: Let g be an evaluation procedure, defined by $r = g(L, x)$ where r is the ranking position of x in the list L.

As more than one target category can be relevant to a query, we also designed a second evaluation setting. In this new procedure, for each query, we retrieved the first 50 results using the method presented in Section 4. to be classified as one of the following judgements: *not relevant*, *relevant* and *highly relevant*, being the last class exclusively used to define the original target category that originated the paraphrase. Each result set was assessed by two fluent or native judges who were instructed to point as relevant those categories that he or she would be interested when searching for the given query. After the judgement process, more than 23,000 categories were analysed. Table 3 shows the inter-rater agreement for each language.

The task allows two evaluation settings. In the first, only the *highly relevant* category is accepted as a positive retrieval, assessing a system strictly by guided paraphrasing as formalised earlier. In the second, categories classified as *relevant* are also considered, allowing a broader evaluation.

Lang.	*Recall*		*MRR*	
	Top 10	*Top 20*	*Top 10*	*Top 20*
EN	0.2922	0.3287	0.1549	0.1575
PT	0.4736	0.5219	0.3151	0.3183
DE	0.7976	0.8392	0.5960	0.5988

Table 4: Recall and MRR results considering only the *highly relevant* categories.

4. Baselines and Results

To provide comparative baselines we implemented a method based on *distributional semantics* to cope with the search of entity categories, whose vectors were generated from unstructured text corpora.

Distributional semantics is based on the hypothesis that words co-occurring in similar contexts tend to have similar meaning (Harris, 1954). Distributional semantics provides representations of the meaning of words in a high-dimensional vector space, which is generated by analysing large-scale text corpora (Turney and Pantel, 2010). The simplification of the meaning representation model supports computation of semantic similarity between two terms by calculating the cosine similarity of their vectors. The baselines applied the Skip-gram (Mikolov et al., 2013) vector space model generated from the Wikipedia 2014 corpora. We pre-processed the corpora lower-casing and stemming each token using the Porter Algorithm (Porter, 1997) and generated the distributional model using the default parameters.

Our baseline is the *sum-algebraic-based method*, where entity categories are compared by an algebraic operation that sums up the vector's components using the resulting vector to calculate the cosine similarity. We developed the experiments with the support of Indra (a distributional semantics tool) (Sales et al., 2018b; Freitas et al., 2016).

The evaluation is applied in two scenarios. The first considers the Top-10 results of each execution and the second considers the Top-20. This assumption makes *precision* a redundant indicator since it can be derived from *recall*. So, the analysis measures recall and mean reciprocal rank (MRR). This methodology of evaluation follows the same strategy used in (Sales et al., 2016) and (Sales et al., 2018a).

Table 4 shows the recall and MRR considering only the *highly relevant* categories. This is the preferable experimental setting, since it evaluates the ability to identify the paraphrases proposed by the volunteers, ignoring any other potential relevant result.

Query Set	Size (# of Queries)
Existing Query Sets	
INEX-XER (Demartini et al., 2010)	55
TREC Entity (Balog et al., 2009)	17
SemSearch ES (Blanco et al., 2011; Halpin et al., 2010)	130
SemSearch LS (Blanco et al., 2011)	43
QALD-2 (Lopez et al., 2013)	140
INEX-LD (Wang et al., 2012)	100
Proposed Query Sets	
Entity Categories English (EN)	220
Entity Categories Portuguese (PT)	229
Entity Categories German (DE)	169

Table 5: Query sets and their sizes in number of queries.

The results show that the proposed baseline performs significantly different across languages. Considering the Top-20 scenario, while the algorithm retrieves about 32% of the English paraphrases, in the German dataset, this rate jumps to more than 82%. This occurs because many of the German paraphrases were described using words that share the same linguistic root of the original category. Furthermore, the English target dataset is significantly larger (345,000 against 235,000) which increases the probability of returning false-positive categories. For the Portuguese dataset, despite the paraphrases were constructed using a rich vocabulary, its target dataset is less than one-third of the English dataset (105,000), which put their results consistently between the other two languages.

Table 3 shows the evaluation considering all relevant categories. In this experiment, in addition to the original target category, we also consider as relevant those categories appointed by both judges simultaneously.

This new scenario of evaluation significantly differs from the previous, specially considering the German language. As we increase the set of relevant categories, the queries do not have the same linguistic roots as before, and so the recall decreases significantly. As both the English and Portuguese languages had already rich paraphrases, their recall did not experienced the same gap. The still higher retrieval score of the Portuguese language is again expressed for its smaller target dataset. Both scenarios of evaluation consistently points the German, Portuguese and English in better positions as shown by the MRR measure.

The dataset and the source code are respectively available at `https://rebrand.ly/cat-paraphases` and `https://rebrand.ly/cat-source-code`.

5. Related Work

Searching of entity categories was not exploited previously as a formatted task. Currently, the most related task is the entity search, which is commonly evaluated using the five query types generated in the context of challenges and campaigns.

Balog and Neumayer (2013) grouped those query sets, normalising their results to point to DBpedia instances.

The XML Entity ranking track (XER) discusses the standardisation of the evaluation procedures for entity retrieval and provides a large dataset sample in which the Wikipedia is used as an underlying collection (Demartini et al., 2010). The track explores two main tasks: Entity Ranking (ER) and Entity List Completion (LC), both of them using the Wikipedia 2009 XML data.

Blanco et al. (2011) present the entity search over Linked Data with keyword queries related to entities or their description. In comparison to TREC 2010 and INEX-XER tracks searches over structured data in RDF rather than unstructured data and XML as a data format, Respectively.

The goal of INEX-LD (Wang et al., 2012) was to investigate retrieval techniques over a combination of textual (wikipedia) and structured data (RDF). Lopez et al. (2013) aimed to present the faults and failures of question answering systems as interfaces to query linked data sources. The SemSearch ES challenge (Blanco et al., 2011; Halpin et al., 2010) in comparison searches over structured data in RDF rather than unstructured data and XML as the previous tasks.

Although these works deal with the entity searching studies, their focus rely on issues of evaluation and shortcomings of different interface to query systems and not aim the study on investigating entity categories. Table 5 shows the size of each query set, along with the size of the query sets proposed in our work.

Sales et al. (2016) proposed a compositional-distributional semantic model to search English entity categories whose syntactic structure plays a fundamental role in composing partial semantic relatedness scores. The method identifies the core concept behind the category and use it to guide the search. Their results, however, cannot be directly compared, since Sales et al. (2016) explores a different dataset.

6. Summary

Despite the high popularity of entity search, entity categories have not been receiving similar attention. In this paper, we shed some light on entitty category descriptors (complex nominals) by presenting the task of semantic search of entity categories and by making it publicly available. The test collections cover three different languages (English, Portuguese and German) and includes baselines. The test collections were designed to meet the demands of the entity search community in providing *more representative* and *long queries* and also by *integrating both structured and unstructured information about entities* (Balog and Neumayer, 2013).

The analysis of the semantic phenomena associated with interpreting natural language entity category allows a focused understanding of how humans define complex predicates and how systems can address complex compositions of predicates. Additionally, the ability to semantically interpret categories and to cope with its associated meaning variations plays a fundamental role in different semantic tasks including *entity search*, *schema-agnostic queries*, *question answering systems* and *text entailment* (Chen et al., 2016; Sales et al., 2016; Freitas, 2015).

However, this task has not been sufficiently individuated, becoming implicit in all of these tasks and not receiving the adequate focus. As a consequence, most discussion on how to cope with semantic variation of categories has been limited in the literature. The semantic search of entity categories can define a new bridge between unstructured and structured data.

7. Acknowledgment

This publication has emanated from research partially supported by the National Council for Scientific and Technological Development, Brazil (CNPq).

8. References

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z., (2007). *DBpedia: A Nucleus for a Web of Open Data*, pages 722–735. Springer, Berlin, Heidelberg.

Balog, K. and Neumayer, R. (2013). A test collection for entity search in DBpedia. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 737–740, New York, USA. ACM.

Balog, K., de Vries, A. P., Serdyukov, P., Thomas, P., and Westerveld, T. (2009). Overview of the TREC 2009 entity track. In *TREC 2009 Working Notes*. NIST, November.

Blanco, R., Halpin, H., Herzig, D. M., Mika, P., Pound, J., Thompson, H. S., and Tran, D. T. (2011). Entity search evaluation over structured web data. In *Proceedings of the 1st International Workshop on Entity-Oriented Search*, Beijing, China, July.

Chen, J., Xiong, C., and Callan, J. (2016). An empirical study of learning to rank for entity search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 737–740, New York, NY, USA. ACM.

Demartini, G., Iofciu, T., and De Vries, A. P. (2010). Overview of the INEX 2009 entity ranking track. In *Proceedings of the Focused Retrieval and Evaluation, and 8th International Conference on Initiative for the Evaluation of XML Retrieval*, INEX'09, pages 254–264, Berlin, Heidelberg. Springer-Verlag.

Elbedweihy, K. M., Wrigley, S. N., Clough, P., and Ciravegna, F. (2015). An overview of semantic search evaluation initiatives. *Web Semantics: Science, Services and Agents on the World Wide Web*, 30(0).

Freitas, A., Barzegar, S., Sales, J. E., Handschuh, S., and Davis, B., (2016). *Semantic Relatedness for All (Languages): A Comparative Analysis of Multilingual Semantic Relatedness Using Machine Translation*, pages 212–222. Springer International Publishing, Cham.

Freitas, A. (2015). *Schema-agnositc queries over large-schema databases: a distributional semantics approach.* Ph.D. thesis, Digital Enterprise Research Institute (DERI), National University of Ireland, Galway.

Halpin, H., Herzig, D. M., Mika, P., Blanco, R., Pound, J., Thompson, H. S., and Tran, D. T. (2010). Evaluating ad-hoc object retrieval. In *Proceedings of the International Workshop on Evaluation of Semantic Technologies*, Shanghai, China, November.

Harris, Z. (1954). Distributional structure. *Word*, 10(23):146–162.

Karen Spärck Jones, editor. (1981). *Information Retrieval Experiment*. Butterworths.

Lopez, V., Unger, C., Cimiano, P., and Motta, E. (2013). Evaluating question answering over linked data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 21:3 – 13. Special Issue on Evaluation of Semantic Technologies.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *ICLR Workshop Papers*.

Porter, M. F. (1997). Readings in information retrieval. chapter An Algorithm for Suffix Stripping, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Pound, J., Mika, P., and Zaragoza, H. (2010). Ad-hoc object retrieval in the web of data. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 771–780, New York, NY, USA. ACM.

Sales, J. E., Freitas, A., Davis, B., and Handschuh, S. (2016). A compositional-distributional semantic model for searching complex entity categories. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics, *SEM@ACL 2016, Berlin, Germany, 11-12 August 2016*.

Sales, J. E., Freitas, A., and Handschuh, S. (2018a). An open vocabulary semantic parser for end-user programming using natural language. In *2018 IEEE Twelfth International Conference on Semantic Computing (ICSC)*, Jan.

Sales, J. E., Souza, L., Barzegar, S., Davis, B., Freitas, A., and Handschuh, S. (2018b). Indra: A word embedding and semantic relatedness server. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 697–706, New York, NY, USA. ACM.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *J. Artif. Int. Res.*, 37(1):141–188, January.

Wang, Q., Kamps, J., Camps, G. R., Marx, M., Schuth, A., Theobald, M., Gurajada, S., and Mishra, A. (2012).

Overview of the INEX 2012 linked data track. In *CLEF 2012 Evaluation Labs and Workshop*, pages 1–13, Rome, Italy.