Linse: A Distributional Semantics Entity Search Engine

Juliano Efson Sales[1], André Freitas[2], Siegfried Handschuh[2], Brian Davis[1]
[1]Insight Centre for Data Analytics
National University of Ireland, Galway
[2]Department of Computer Science and Mathematics
University of Passau

ABSTRACT

Entering *Football Players from United States* when searching for *American Footballers* is an example of *vocabulary mismatch*, which occurs when different words are used to express the same concepts. In order to address this phenomenon for entity search targeting descriptors for *complex categories*, we propose a *compositional-distributional semantics* entity search engine, which extracts semantic and commonsense knowledge from large-scale corpora to address the vocabulary gap between query and data.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

semantic search, entity search, distributional search.

1. INTRODUCTION

Entities naming *sets* and *categories* are fundamental for describing structured, semi-structured and unstructured data and are present from Wikis to databases. The natural language descriptors associated with these entities are fundamental to support users searching, querying or browsing structured or semi-structured data, which heavily depend on them to find the desired piece of information.

Users searching over a set of *natural language category descriptors* (NLCDs) demand principled mechanisms to support crossing the semantic gap between the user queries and the target NLCDs. The *vocabulary problem* occurs when different terms are used to express the same concepts. One of the first studies in this direction shows that 80% of people familiar with the same domain use different terms to name the same concept[4]. Search over collections which do not have high vocabulary redundancy is highly affected by the vocabulary problem, since users do not know which words were used to describe target entities, categories or documents.

Shifting from unique words to multi-word expressions enables differences not only in vocabulary, but also in ordering and *compositionality*, increasing the complexity in addressing the vocabulary mismatch. For example, *Football Players from United States* would be an acceptable paraphrase for *American Footballers*.

In order to address this problem, we have developed *Linse*, an entity search engine which implements a *compositional-distributional semantic model* to address the vocabulary problem for searching entities which describe complex categories (e.g. 'Buildings destroyed in the great fire of London and not rebuilt'). This paper demonstrates the basic mechanisms and principles behind the Linse entity search engine, showing how distributional semantics can support users in crossing the query-data semantic gap.

2. THE COMPOSITIONAL-DISTRIBUTIONAL SEMANTIC MODEL

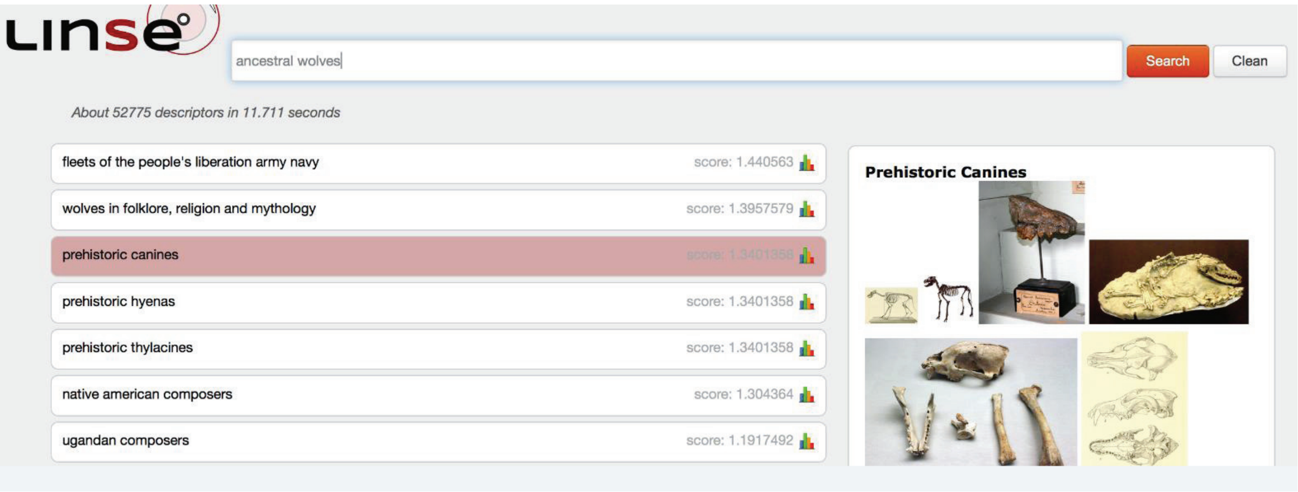
Linse uses distributional semantic models [3] (semantic models which are automatically built from large-scale corpora) as a principled mechanism to address the vocabulary gap between users and data. On the top of the distributional model, a compositional model based on a *semantic pivoting mechanism* [5] is used to provide a semantic interpretation of the query under the Knowledge Base (KB) of category descriptors (NLCDs).

The Linse search engine decomposes every NLCD into a *semantic core* and three types of *specializations*: (i) specializations of the core defined by noun, adjective or adverb modifiers (e.g. 'Tall Athletes'); (ii) specializations defined by named entities ('Athletes from Russia') ; (iii) specializations defined by temporal references ('Athletes born in the 80s').

The Linse approach determines the *constituency structure* of every NLCD during indexing time, generating a graph containing a semantic core and its specializations, following the semantic representation model for categories described in [5]. Each core and specialization type is indexed in a different distributional vector space, which maps to different indexes.

The *query interpretation process* starts with the syntactic parsing of the query. The noun phrase head defines the query semantic core. For the example query 'ancestral wolves', the semantic core is 'wolves'. The semantic core is sent as a query to the *semantic core distributional vector space* which computes the distributional semantic re-

SIGIR '15, August 09-13, 2015, Santiago, Chile
ACM 978-1-4503-3621-5/15/08.
http://dx.doi.org/10.1145/2766462.2767871



ancestral wolves

About 52775 descriptors in 11.711 seconds

fleets of the people's liberation army navy score: 1.440563
wolves in folklore, religion and mythology score: 1.3957579
prehistoric canines score: 1.3401358
prehistoric hyenas score: 1.3401358
prehistoric thylacines score: 1.3401358
native american composers score: 1.304364
ugandan composers score: 1.1917492

Figure 1: Screenshot of the Linse search result.

latedness between the query term and the indexed semantic cores, returning a set of most related semantic cores. For the example query, the corresponding 'canine' semantic core is returned.

The next step consists in selecting the next core modifier, the adjective 'ancestral', sending it as a query to the subspace of specializations of 'canine'. The use of 'canine' as a *semantic pivot* supports the reduction of the dimensionality of the distributional space, by only looking into specializations which are applied to 'canine'. Using the distributional semantic relatedness measure over the specializations subspace, it returns 'prehistoric canine' as an output. The ranking approach uses as a ranking score a weighted composition of the distributional relatedness measure between each query term and the matching NLCD terms. In the context of this work, Explicit Semantic Analysis [6] is used as a semantic relatedness measure.

Figure 1 shows an screenshot of the result set for the example query, which displays the search box, the ranked list of returned categories and their associated ranking scores. The use of the distributional semantic relatedness as a ranking function provides a comprehensive semantic matching mechanism [5] which is suitable for exploratory entity search scenarios.

For the demonstration, a knowledge base of more than 300,000 category descriptors corresponding to the Wikipedia category links were indexed. A set of natural language queries over the knowledge base are demonstrated in an online video[1]. Some of these queries with their corresponding NLCD mappings are listed on Table 1.

Linse was implemented using Java over Lucene and used the EasyESA framework [2] over the Wikipedia 2013 corpus.

3. SUMMARY

This paper demonstrated the Linse entity search engine. Linse uses a compositional-distributional model to provide a mechanism to search for complex category descriptors. The use of distributional semantic models provides a low-effort (automatically built from large-scale corpora) and comprehensive semantic matching strategy, which can be easily transported to other languages. The proposed model can be applied in different scenarios for entity search, varying from tag recommendation systems to Question Answering over Linked Data.

Queries → NLCDs
literature journals from austria → austrian literary magazines
movie premieres in 2008 → 2008 film festivals
obsolete justice institutions of connecticut → defunct law enforcement agencies of connecticut
norwegian top singles → number-one singles in norway
italian social movements → political movements in italy
repeating events which were initiated in 1875 → recurring events established in 1875
horse enthusiasts from france → french racehorse owners and breeders

Table 1: Examples of query-NLCD mappings.

4. ACKNOWLEDGMENTS

This publication was supported by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

5. REFERENCES

[1] A. Freitas, et al., On the Semantic Representation and Extraction of Complex Category Descriptors, In Proc. of NLDB, 2014.

[2] D. Carvalho, et al., EasyESA: A Low-effort Infrastructure for Explicit Semantic Analysis, In Proc. of ISWC, 2014.

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

[4] G. W. Furnas et al., The vocabulary problem in human-system communication. Commun. ACM, 30(11): pp. 964 - 971, November 1987.

[5] A. Freitas, E. Curry, Natural Language Queries over Heterogeneous Linked Data Graphs, In Proc. of IUI, 2014.

[6] E. Gabrilovich, S. Markovitch, Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In Proc. of IJCAI, 2007.

[1] http://treo.deri.org/linsedemo