BUT-TYPED: Using domain knowledge for computing typed similarity

Lubomir Otrusina
Brno University of Technology
Faculty of Information Technology
IT4Innovations Centre of Excellence
Bozetechova 2, 612 66 Brno
Czech Republic
iotrusina@fit.vutbr.cz

Pavel Smrz
Brno University of Technology
Faculty of Information Technology
IT4Innovations Centre of Excellence
Bozetechova 2, 612 66 Brno
Czech Republic
smrz@fit.vutbr.cz

Abstract

This paper deals with knowledge-based text processing which aims at an intuitive notion of textual similarity. Entities and relations relevant for a particular domain are identified and disambiguated by means of semi-supervised machine learning techniques and resulting annotations are applied for computing typed-similarity of individual texts.

The work described in this paper particularly shows effects of the mentioned processes in the context of the *SEM 2013 pilot task on typed-similarity, a part of the Semantic Textual Similarity shared task. The goal is to evaluate the degree of semantic similarity between semi-structured records. As the evaluation dataset has been taken from Europeana – a collection of records on European cultural heritage objects – we focus on computing a semantic distance on field *author* which has the highest potential to benefit from the domain knowledge.

Specific features that are employed in our system BUT-TYPED are briefly introduced together with a discussion on their efficient acquisition. Support Vector Regression is then used to combine the features and to provide a final similarity score. The system ranked third on the attribute *author* among 15 submitted runs in the typed-similarity task.

1 Introduction

The goal of the pilot typed-similarity task lied in measuring a degree of semantic similarity between semi-structured records. The data came from the Europeana digital library[1] collecting millions of records on paintings, books, films, and other museum and archival objects that have been digitized throughout Europe. More than 2,000 cultural and scientific institutions across Europe have contributed to Europeana. There are many metadata fields attached to each item in the library, but only fields *title, subject, description, creator, date* and *source* were used in the task.

Having this collection, it is natural to expect that domain knowledge on relevant cultural heritage entities and their inter-relations will help to measure semantic closeness between particular items. When focusing on similarities in a particular field (a semantic type) that clearly covers a domain-specific aspect (such as field author/creator in our case), the significance of the domain knowledge should be the highest.

Intuitively, the semantic similarity among authors of two artworks corresponds to strengths of links that can be identified among the two (groups of) authors. As the gold standard for the task resulted from a Mechanical Turk experiment (Paolacci et al., 2010), it could be expected that close fields correspond to authors that are well known to represent the same style, worked in the same time or the same art branch (e. g., Gabriël Metsu and Johannes Vermeer), come from the same region (often guessed from the names), dealt with related topics (not necessarily in the artwork described by the record in question), etc. In addition to necessary evaluation of the intersection and the union of two author fields (leading naturally to the Jaccard similarity coeffi-

[1] http://www.europeana.eu/

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*, pages 119–123, Atlanta, Georgia, June 13-14, 2013. ©2013 Association for Computational Linguistics

cient on normalized name records – see below), it is therefore crucial to integrate means measuring the above-mentioned semantic links between identified authors.

Unfortunately, there is a lot of noise in the data used in the task. Since Europeana does not precisely define meaning and purpose of each particular field in the database, many mistakes come directly from the unmanaged importing process realized by participating institutions. Fields often mix content of various semantic nature and, occasionally, they are completely misinterpreted (e. g., field *creator* stands for the author, but, in many cases, it contains only the institution the data comes from). Moreover, the data in records is rather sparse – many fields are left empty even though the information to be filled in is included in original museum records (e. g., the author of an artwork is known but not entered).

The low quality of underlying data can be also responsible for results reported in related studies. For example, Aletras et al. (2012) evaluate semantic similarity between semi-structured items from Europeana. They use several measures including a simple normalized textual overlap, the extended Lesk measure, the cosine similarity, a Wikipedia-based model and the LDA (Latent Dirichlet Allocation). The study, restricted to fields *title, subject* and *description*, shows that the best score is obtained by the normalized overlap applied only to the title field. Any other combination of the fields decreased the performance. Similarly, sophisticated methods did not bring any improvement.

The particular gold standard (training/test data) used in the typed-similarity task is also problematic. For example, it provides estimates of location-based similarity even though it makes no sense for particular two records – no field mentions a location and it cannot be inferred from other parts). A throughout analysis of the task data showed that *creator* is the only field we could reasonably use in our experiments (although many issues discussed in previous paragraphs apply for the field as well). That is why we focus on similarities between author fields in this study.

While a plenty of measures for computing textual similarity have been proposed (Lin, 1998; Landauer et al., 1998; Sahlgren, 2005; Gabrilovich and Markovitch, 2007) and there is an active research in the fields of Textual Entailment (Negri et al., 2012), Paraphrase Identification (Lintean and Rus, 2010) and, recently, the Semantic Textual Similarity (Agirre et al., 2012), the semi-structured record similarity is a relatively new area of research. Even though we focus on a particular domain-specific field in this study, our work builds on previous results (Croce et al., 2012; Annesi et al., 2012) to pre-compute semantic closeness of authors based on available biographies and other related texts.

The rest of the paper is organized as follows: The next section introduces the key domain-knowledge processing step of our system which aims at recognizing and disambiguating entities relevant for the cultural heritage domain. The realized system and its results are described in Section 3. Finally, Section 4 briefly summarizes the achievements.

2 Entity Recognition and Disambiguation

A fundamental step in processing text in particular fields lies in identifying named entities relevant for similarity measuring. There is a need for a named entity recognition tool (NER) which identifies names and classifies referred entities into predefined categories. We take advantage of such a tool developed by our team within the DECIPHER project[2].

The DECIPHER NER is able to recognize artists relevant for the cultural heritage domain and, for most of them, to identify the branch of the arts they were primarily focused on (such as painter, sculptors, etc.). It also recognizes names of artworks, genres, art periods and movements and geographical features. In total, there are 1,880,985 recognizable entities from the art domain and more than 3,000,000 place names. Cultural-heritage entities come from various sources; the most productive ones are given in Table 1. The list of place names is populated from the Geo-Names database[3].

The tool takes lists of entities and constructs a finite state automaton to scan and annotate input texts. It is extremely fast (50,000 words per second) and has a relatively small memory footprint (less than 90 MB for all the data).

Additional information attached to entities is

[2] http://decipher-research.eu/

[3] http://www.geonames.org/

Source	# of entities
Freebase[4]	1,288,192
Getty ULAN[5]	528,921
VADS[6]	31,587
Arthermitage[7]	4,259
Artcyclopedia[8]	3,966

Table 1: Number of art-related entities from various sources

stored in the automaton too. A normalized form of a name and its semantic type is returned for each entity. Normalized forms enable identifying equivalent entities expressed differently in texts, e. g., Gabriël Metsu refers to the same person as Gabriel Metsu, US can stand for the United States (of America), etc. Type-specific information is also stored. It includes a detailed type (e. g., architect, sculptor, etc.), nationality, relevant periods or movements, and years of birth and death for authors. Types of geographical features (city, river), coordinates and the GeoNames database identifiers are stored for locations.

The tool is also able to disambiguate entities based on a textual context in which they appeared. Semantic types and simple rules preferring longer matches provide a primary means for this. For example, a text containing *Bobigny – Pablo Picasso*, refers probably to a station of the Paris Metro and does not necessarily deal with the famous Spanish artist. A higher level of disambiguation takes form of classification engines constructed for every ambiguous name from Wikipedia. A set of most specific terms characterizing each particular entity with a shared name is stored together with an entity identifier and used for disambiguation during the text processing phase. Disambiguation of geographical names is performed in a similar manner.

3 System Description and Results

To compute semantic similarity of two non-empty author fields, normalized textual content is compared by an exact match first. As there is no unified form defined for author names entered to the field, the next step applies the NER tool discussed in the previous section to the field text and tries to identify all mentioned entities. Table 2 shows examples of texts from author fields and their respective annotations (in the typewriter font).

Dates and places of birth and death as well as few specific keywords are put together and used in the following processing separately. To correctly annotate expressions that most probably refer to names of people not covered by the DECIPHER NER tool, we employ the Stanford NER[9] that is trained to identify names based on typical textual contexts.

The final similarity score for a pair of author fields is computed by means of the SVR combining specific features characterizing various aspects of the similarity. Simple Jaccard coefficient on recognized person names, normalized word overlap of the remaining text and its edit distance (to deal with typos) are used as basic features.

Places of births and deaths, author's nationality (e. g., Irish painter) and places of work (active in Spain and France) provide data to estimate location-based similarity of authors. Coordinates of each location are used to compute an average location for the author field. The distance between the average coordinates is then applied as a feature. Since types of locations (city, state, etc.) are also available, the number of unique location types for each item and the overlap between corresponding sets are also employed as features.

Explicitly mentioned dates as well as information provided by the DECIPHER NER are compared too. The time-similarity feature takes into account time overlap of the dates and time distance of an earlier and a later event.

Other features reflect an overlap between visual art branches represented by artists in question (Photographer, Architect, etc.), an overlap between their styles, genres and all other information available from external sources. We also employ a matrix of artistic influences that has been derived from a large collection of domain texts by means of relation extraction methods.

Finally, general relatedness of artists is pre-computed from the above-mentioned collection by means of Random Indexing (RI), Explicit Semantic Analysis (ESA) and Latent Dirichlet Allocation (LDA) methods, stored in sparse matrices and entered as a final set of features to the SVR process.

The system is implemented in Python and takes

[9] http://nlp.stanford.edu/software/CRF-NER.shtml

Eginton, Francis; West, Benjamin `<author name="Francis Eginton" url="http://www.freebase.com/m/0by1w5n"> Eginton, Francis</author>; <author name="Benjamin West" url="http://www.freebase.com/m/01z6r6">West, Benjamin</author>`
Yossef Zaritsky Israeli, born Ukraine, 1891-1985 `<author name="Joseph Zaritsky" url="http://www.freebase.com/m/0bh71xw" nationality="Israel" place_of_birth="Ukraine" date_of_birth="1891" date_of_death="1985">Yossef Zaritsky Israeli, born Ukraine, 1891-1985</author>`
Man Ray (Emmanuel Radnitzky) 1890, Philadelphia – 1976, Paris `<author name="Man Ray" alternate_name="Emmanuel Radnitzky" url="http://www.freebase.com/m/0gskj" date_of_birth="1890" place_of_birth="Philadelphia" date_of_death="1976" place_of_death="Paris"> Man Ray (Emmanuel Radnitzky) 1890, Philadelphia - 1976, Paris</author>`

Table 2: Examples of texts in the author field and their annotations

advantage of several existing modules such as gensim[10] for RI, ESA and other text-representation methods, numpy[11] for Support Vector Regression (SVR) with RBF kernels, PyVowpal[12] for an efficient implementation of the LDA, and nltk[13] for general text pre-processing.

The resulting system was trained and tested on the data provided by the task organizers. The train and test sets consisted each of 750 pairs of cultural heritage records from Europeana along with the gold standard for the training set. The BUT-TYPED system reached score 0.7592 in the author field (cross-validated results, Pearson correlation) on the training set where 80 % were used for training whereas 20 % for testing. The score for the field on the testing set was 0.7468, while the baseline was 0.4278.

4 Conclusions

Despite issues related to the low quality of the gold standard data, the attention paid to the similarity computation on the chosen field showed to bear fruit. The realized system ranked third among 14 others in the criterion we focused on. Domain knowledge proved to significantly help in measuring semantic closeness between authors and the results correspond to an intuitive understanding of the similarity between artists.

Acknowledgments

This work was partially supported by the EC's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 270001, and by the Centrum excellence IT4Innovations (ED1.1.00/02.0070).

References

Agirre, E., Diab, M., Cer, D., and Gonzalez-Agirre, A. (2012). Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 385–393. Association for Computational Linguistics.

Aletras, N., Stevenson, M., and Clough, P. (2012). Computing similarity between items in a digital library of cultural heritage. *Journal on Computing and Cultural Heritage (JOCCH)*, 5(4):16.

Annesi, P., Storch, V., and Basili, R. (2012). Space projections as distributional models for semantic composition. In *Computational Linguistics and Intelligent Text Processing*, pages 323–335. Springer.

[10] http://radimrehurek.com/gensim/

[11] http://www.numpy.org/

[12] https://github.com/shilad/PyVowpal

[13] http://nltk.org/

Croce, D., Annesi, P., Storch, V., and Basili, R. (2012). Unitor: combining semantic text similarity functions through sv regression. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 597–602. Association for Computational Linguistics.

Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 6–12.

Landauer, T., Foltz, P., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2):259–284.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, volume 1, pages 296–304. Citeseer.

Lintean, M. C. and Rus, V. (2010). Paraphrase identification using weighted dependencies and word semantics. *Informatica: An International Journal of Computing and Informatics*, 34(1):19–28.

Negri, M., Marchetti, A., Mehdad, Y., Bentivogli, L., and Giampiccolo, D. (2012). semeval-2012 task 8: Cross-lingual textual entailment for content synchronization. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 399–407. Association for Computational Linguistics.

Paolacci, G., Chandler, J., and Ipeirotis, P. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making*, 5(5):411–419.

Sahlgren, M. (2005). An introduction to random indexing. In *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005*. Citeseer.