Computing Text Semantic Relatedness using the Contents and Links of a Hypertext Encyclopedia: Extended Abstract*

Majid Yazdani
Idiap Research Institute and EPFL
Centre du Parc, Rue Marconi 19
1920 Martigny, Switzerland
majid.yazdani@idiap/epfl.ch

Andrei Popescu-Belis
Idiap Research Institute
Centre du Parc, Rue Marconi 19
1920 Martigny, Switzerland
andrei.popescu-belis@idiap.ch

Abstract

We propose methods for computing semantic relatedness between words or texts by using knowledge from hypertext encyclopedias such as Wikipedia. A network of concepts is built by filtering the encyclopedia's articles, each concept corresponding to an article. A random walk model based on the notion of Visiting Probability (*VP*) is employed to compute the distance between nodes, and then between sets of nodes. To transfer learning from the network of concepts to text analysis tasks, we develop two common representation approaches. In the first approach, the shared representation space is the set of concepts in the network and every text is represented in this space. In the second approach, a latent space is used as the shared representation, and a transformation from words to the latent space is trained over *VP* scores. We applied our methods to four important tasks in natural language processing: word similarity, document similarity, document clustering and classification, and ranking in information retrieval. The performance is state-of-the-art or close to it for each task, thus demonstrating the generality of the proposed knowledge resource and the associated methods.

1 Introduction

Estimating the semantic relatedness of two text fragments – such as words, sentences, or entire documents – is important for many natural language processing or information retrieval applications, such as word sense disambiguation, coreference resolution, information extraction patterns, or semantic indexing. Existing measures of semantic relatedness based on lexical overlap, though widely used, are of little help when text similarity is not based on identical words, while statistically-based topic models, such as PLSA or LDA, do not make use of structured knowledge, now available on a large scale, to go beyond word distribution properties.

In this extended abstract of our paper published in *Artificial Intelligence* [Yazdani and Popescu-Belis, 2013], we show how to compute semantic relatedness between sets of words using the knowledge enclosed in a large hypertext encyclopedia (e.g., the English Wikipedia). We first build a network of concepts from encyclopedic articles, with two types of links between them (Section 2). Then, we define a proximity measure between sets of concepts (Section 3) and present two ways to build a shared representation space for texts (Section 4). Finally, we briefly review (in Section 5) some results obtained on four tasks from natural language processing and information retrieval – all discussed in detail in the journal paper. The results demonstrate that our method brings a unified and robust solution to measuring semantic relatedness.

2 Wikipedia as a Network of Concepts

We built our concept network from Wikipedia by using the Freebase Wikipedia Extraction (WEX) dataset [Metaweb Technologies, 2010] (version dated 2009-06-16). Not all Wikipedia articles were considered appropriate to include in the network of concepts, for reasons related to their nature and reliability, but also to the tractability of the overall method, given the very large number of pages in the English Wikipedia. Therefore, we removed all Wikipedia articles that belonged to the following name spaces: Talk, File, Image, Template, Category, Portal, and List, because these articles do not describe concepts, but contain auxiliary media and information that do not belong into the concept network. Also, disambiguation pages were removed as well, as they only point to different meanings of the title or of its variants. Moreover, articles with less than 100 non-stop words are removed from the final set yielding a resulting set of 1,264,611 concepts.

We consider two types of links between concepts derived from the hyperlinks and content of articles in Wikipedia – for more types, see [Yazdani and Popescu-Belis, 2010]. The first type of links are the *hyperlinks between articles*. The use of hyperlinks embodies the somewhat evident observation that every hyperlink from the content of an article towards another one indicates a certain relation between the two articles. These are encyclopedic or pragmatic relations, i.e. between concepts in the world, and subsume semantic relatedness. In other words, if article A contains a hyperlink towards article B, then B helps to understand A, and B is considered to be related to A.

The second type of links is based on *similarity of lexical*

*This paper is an extended abstract of the AI Journal publication [Yazdani and Popescu-Belis, 2013].

content between articles of Wikipedia, computed from word co-occurrence. If two articles have many words in common, then a topic-similarity relation holds between them. To capture content similarity, we computed the lexical similarity between articles as the cosine similarity between the vectors derived from the articles' texts, after stopword removal and stemming. We then linked every article to its k most similar articles, with a weight according to the normalized lexical similarity score. As the Wikipedia articles are scattered in the space of words, tuning k does not seem to bring crucial changes. If k is very small then the neighborhood contains little information, whereas a large k makes computation time-consuming. Typically, $k = 10$ in our experiments.

3 Aggregated Proximity Measure

Our goal is first to estimate a distance between two nodes in a network by taking into account the global connectivity of the network, and without being biased by local properties. Indeed, the use of individual links and paths, e.g. when estimating proximity as the length of shortest path, does not take into account their relative importance with respect to the overall properties of the network, such as the number and length of all possible paths between two nodes. Moreover, the length of the shortest path is quite sensitive to spurious links. Therefore, a number of aggregated proximity measures based on random walk have been proposed in the literature, such as PageRank (including Personalized PageRank) and hitting time. Previous studied showed that these aggregated measures are more effective than individual links and paths [Brand, 2005], [Sarkar and Moore, 2007], [Liben-Nowell and Kleinberg, 2003].

Following a similar motivation, and using a random walk approach, we define the proximity of two nodes as the *Visiting Probability (VP)* of a random walker going from one node to the other one.

If A_l is the weighted adjacency matrix of link type l ($1 \leq l \leq L$), then the terms of the transition matrix C_l that gives the probability of a direct (one step) transition between nodes i and j using only links of type l can be written as $C_l(i,j) = A_l(i,j) / \sum_{k=1}^{n} A_l(i,k)$. In the random walk process using all link types ($1 \leq l \leq L$), if the weight w_l ($\sum_l w_l = 1$) is the importance of link type l, then the overall transition matrix C which gives the transition probability $C_{i,j}$ between any nodes i and j is $C = \sum_{l=1}^{L} w_l C_l$.

Given the nodes i and j in the network, VP_{ij} is the probability of visiting j for the *first time* when a random walker starts from i in the network. We introduce C' as being equal to the transition matrix C, except that in row j, $C'(j,k) = 0$ for all k. This indicates the fact that when the random walker visits j for the first time, it can not exit from it and its probability mass drops to zero in the next step. This modified transition matrix was defined to account for the definition of *VP* as the probability of the *first* visit of j.

Finally, the *VP* from i to j can be formulated recursively as follows: $VP_{ij}^{t} = \alpha \times \sum_k C'(i,k) VP_{kj}^{t-1}$ with $VP_{ij}^{0} = 0$, $VP_{jj}^{t} = 1$ and α is a dampening parameter.

In the journal paper [Yazdani and Popescu-Belis, 2013], we showed that *VP* reduces the effect of spurious links and of popular pages in a network. Moreover, the definition of *VP* allowed us to design fast approximation algorithms applicable to large networks.

4 Common Representation Space for Transfer Learning

In order to transfer human knowledge embodied in the Wikipedia network of concepts towards a measure of text similarity in various text analysis tasks, we need to build a shared representation for text fragments. We propose two shared representation models (i.e. spaces), which we explain in the following sections below.

The network of concepts built from Wikipedia consists of many concepts from human knowledge. Therefore, the set of concepts in the network is rich enough so that it can represent the content of any text fragment. The first shared representation we develop is the set of concepts in the network: a given text is simply mapped to the corresponding concepts in this network. Then, to compute similarity between two texts, *VP* similarity is applied to compute the distance between the two sets of nodes (concepts).

The second method uses the latent space model that we explain in Section 4.2 as the shared representation. In this approach, we assume that there is a latent space in which semantically similar texts are placed in close positions and semantically unrelated texts are placed farther away from each other. We showed that each concept in the network (corresponding to a Wikipedia article) has a text body which explains the concept. We learn a transformation from words in the title and body to the latent space so that two similar concepts in terms of *VP* are in close distance. Therefore, to transfer knowledge from the network to any processing method that uses feature vectors, texts are transformed using the learned transformation into the latent space.

4.1 Mapping Text Fragments to Concepts in the Network

For mapping, two cases must be considered, according to whether the text matches exactly the title of a Wikipedia page or not. Exact matching is likely to occur with individual words or short phrases, but not with entire sentences or longer texts.

If a text fragment consists of a single word or a phrase that *matches exactly the title of a Wikipedia page*, then it is simply mapped to that concept. In the case of words or phrases that may refer to several concepts in Wikipedia, we simply assign to them the same page as the one assigned by the Wikipedia contributors as the most salient or preferred sense or denotation. For instance, 'mouse' directs to the page about the animal, which contains an indication that the 'mouse_(computing)' page describes the pointing device, and that other senses are listed on the 'mouse_(disambiguation)' page. So, here, we simply map 'mouse' to the animal concept. However, for other words, no sense or denotation is preferred by the Wikipedia contributors, e.g. for the word 'plate'. In such cases, a disambiguation page is associated to that word or phrase. We chose not to include such pages

in our network, as they do not correspond to individual concepts. So, in order to select the referent page for such words, we simply use the lexical similarity approach we will now describe.

When a fragment (a word, phrase, sentence, or text) *does not match exactly the Wikipedia title of a vertex in our network*, it is mapped to the network by computing its lexical similarity with the text content of the vertices in the network, using cosine distance over stemmed words, stopwords being removed. The text fragment is mapped to the k most similar articles according to this similarity score, resulting in a set of at most k weighted concepts. The weights are normalized, summing up to one, therefore the text representation in the network is a *probability distribution over at most k concepts*.

This mapping algorithm has an important role in the performance of the final system, in combination with the network distance (*VP*). It must however be noted that the effects of wrong mappings at this stage are countered later on. For instance, when large sets of concepts related to two text fragments are compared, a few individual mistakes are not likely to alter the overall relatedness scores. Alternatively, when comparing individual words, wrong mappings are less likely to occur because the test sets for word similarity described in [Rubenstein and Goodenough, 1965], [Miller and Charles, 1991], and by [Finkelstein *et al.*, 2002] also consider implicitly the most salient sense of each word, just as described above for Wikipedia.

4.2 Learning Embeddings to Latent Space

The second method we propose to measure text similarity using *VP* is to learn an embedding (transformation) from words to a latent space using as a criterion the *VP* scores on the Wikipedia concept network. Learning latent space model over preference data have been studied previously in [Grangier and Bengio, 2008; Weston *et al.*, 2011; Bai *et al.*, 2010], we follow the same main direction here.

At training time, given a series of samples – that is, pairs of texts with *VP* values from the first text to the second one – the goal is to learn a transformation from the space of words to a latent space, so that the similarity between the latent representation of the texts is as close as possible to the *VP* similarity. In other words, the goal is to approximate *VP* between two texts i and j by the matrix product $x_i A B' x'_j$, where x_i and x_j are the TF-IDF vectors of the two texts constructed from their words using a fixed dictionary. The size of matrices A and B is $n \times m$, with n being the size of the dictionary (number of words) and m the size of the latent space (akin to the number of topics in topic models). Two different matrices A and B are needed because *VP* values are not symmetric in i and j.

In principle, all pairs of Wikipedia articles (i.e., texts) corresponding to nodes in our network can be used for training, but this set is extremely large (ca. 1.4×10^{12}) and moreover, we showed that the most values are close to zero and are not valuable for training. Therefore, we formulate the following constraints for training: (1) training should focus on neighboring articles (articles with high *VP* values), and (2) the exact values of *VP* are replaced with the ranking of pairs of articles by decreasing *VP*. We show here that under these constraints valuable embeddings can be learned.

Let $VPto_k(i)$ be the set of the k closest articles *to* the article i according to *VP* similarity. We define a hinge loss function L as follows, so that the similarity between i and its k closest articles is larger than the similarity to all other articles by a fixed margin M.

$$L = \sum_{i \in \text{WP}} \sum_{j \in VPto_k(i)} \sum_{z \notin VPto_k(i)} max(0, M - x_i A B' x'_j + x_i A B' x'_z)$$

We optimize L with stochastic gradient descent: in each iteration we randomly choose one article i, then randomly choose one of the k closest articles to i (noted j) and one other article from the rest of documents (noted z).

Moreover, to perform regularization over matrices A and B when optimizing L, we impose the constraint that A and B are orthonormal. In order to apply this constraint, we project at every 1000 iterations both A and B to their nearest orthogonal matrix found by using SVD decomposition. The rationale for the constraint is the following: if we assume that each latent dimension corresponds to a possible topic or theme, then these should be as orthogonal as possible.

The two main findings from training the embeddings are: First, *VP* over the hyperlinks graph is harder to learn, which may be due to the fact that hyperlinks are defined by users in a manner that is not totally predictable. Second, regularization decreases the prediction ability. However, if regularization traded prediction power for more generality, in other words if it reduced overfitting to this problem and made the distance more general, then it would still constitute a useful operation. This is checked in the experiments in [Yazdani and Popescu-Belis, 2013].

Learning embeddings in comparison to mapping to the concept network has some advantages. The main one is that it can be applied with a much lower cost at run time and make it independent of graph size. Moreover, it can be more easily integrated as prior knowledge to other learning algorithms for NLP, and it can be applied to very large scale problems.

4.3 Properties of the Resulting Network

The processing of the English Wikipedia resulted in a very large network of concepts. The network has more than 1.2 million nodes (i.e. vertices), with an average of 28 outgoing hyperlinks per node and 10 outgoing content links per node.

A natural question arising at this point is: how can the structure of the network be characterized, apart from putting it to work? A number of quantitative parameters have been proposed in graph theory and social network analysis, and some have for instance been used to analyze WordNet (and an enhanced version of it) by [Navigli and Lapata, 2010]. We compute below some well-known parameters for our network, and add a new, more informative characterization.

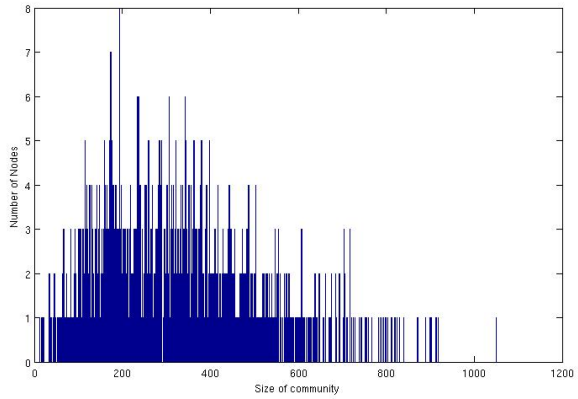
A first characteristic of graphs is their degree distribution, which for the original Wikipedia with hyperlinks

seems to follow a power law. A more relevant property here is the network clustering coefficient, which is the average of clustering coefficients per node, defined as the size of the immediate neighborhood of the node divided by the maximum number of links that could connect all pairs of neighbors [Watts and Strogatz, 1998].

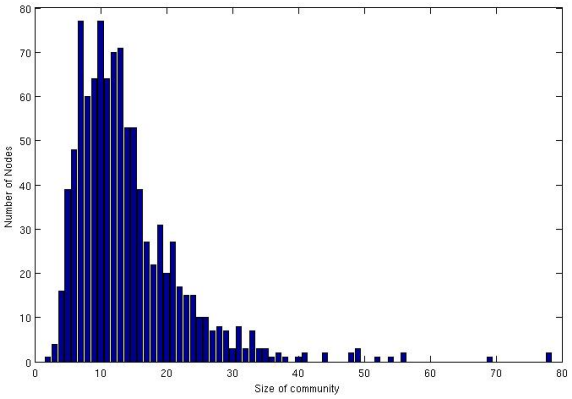
For our hyperlink graph, the value of this coefficient is 0.16, while for the content link graph it is 0.26. These values show that the hyperlink graph is less clustered than the content link one, i.e. the distribution of nodes and links is more homogeneous, and that overall the two graphs are rather weakly clustered. The observed values, together with the power law degree distribution, suggest that our graph is a scale-free network – characterized by the presence of "hub" nodes – or a small-world network [Watts and Strogatz, 1998].

Moreover, an ad-hoc measure offers an even better illustration of the network's topology. Its goal is to measure how much the graph is clustered, i.e. whether communities of nodes based on neighborhoods have a preferred size, or are uniformly distributed. We consider a sample of 1000 nodes, and for each node of the sample, the Personalized PageRank algorithm [Haveliwala, 2003] is initialized from it. This results in a proximity coefficient for each node in the graph to the initial node. The community size for the initial node is computed by sorting all nodes with respect to their proximity and counting how many nodes contribute to 99% of the mass. A barplot of these values, sorted by community size, is shown respectively for hyperlinks and for content links in Figures 1 (a) and (b).

The values shown in Figure 1 show that the distribution is neither flat nor uniformly decreasing, but has a peak, which provides an indication of the average size of clusters. This size is around 150–400 nodes for the hyperlink graph, without a sharp maximum, showing less clustering than for content links, for which this average is around 7–14 nodes. The use of hyperlinks thus avoids local clusters and extends considerably the connectivity of the network in comparison to content similarity ones.



(a) Hyperlink graph



(b) Content link graph

Figure 1: Distribution of community sizes for a sample of 1000 nodes. For each community size (x-axis) the graphs show the number of nodes (y-axis) having a community of that size. Both graphs have a tendency towards clustering, but with different average sizes.

5 Overview of Experimental Results

To evaluate the proposed distance, we applied our method to four important tasks in natural language processing: word similarity, document similarity, document clustering, and document classification, along with unsupervised information retrieval and learning to rank [Yazdani and Popescu-Belis, 2013]. The performance of our method is state-of-the-art or close to it for all the tasks, thus demonstrating the generality of the method and the utility of the accompanying knowledge resource. Moreover, we show that using both hyperlinks and lexical similarity links improves the scores with respect to a method using only one of them, because hyperlinks bring additional real-world knowledge not captured by lexical similarity.

Moreover, the embeddings learned on *VP* similarities achieve competitive results on the data sets, while requiring a much shorter computation time at the query stage (testing). The regularization imposed on the embeddings reduced their predictive power for the *VP* similarities, but we showed that it improved the performance on the text similarity tasks.

A distance-based classifier and ranker was designed and trained using the embeddings as the initial state of the distance metric. The resulting classifier was tested on text classification and information retrieval tasks. The main observation was that, when the training set is small, the distance learning algorithm initialized with the embeddings from *VP* similarities over Wikipedia graphs outperformed the baseline algorithm significantly. By adding more and more labeled data, the importance of prior knowledge appears to decrease, because the distance learning algorithm can infer reliable decisions based only on the training data.

Acknowledgments

This work has been supported by the Swiss National Science Foundation through the National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM2), http://www.im2.ch.

References

[Bai *et al.*, 2010] Bing Bai, Jason Weston, David Grangier, Ronan Collobert, Kunihiko Sadamasa, Yanjun Qi, Olivier Chapelle, and Kilian Weinberger. Learning to rank with

(a lot of) word features. *Information Retrieval*, 13(3):291–314, 2010.

[Brand, 2005] Matthew Brand. A random walks perspective on maximizing satisfaction and profit. In *Proceedings of SDM 2005 (SIAM International Conference on Data Mining)*, pages 12–19, Newport Beach, CA, 2005.

[Finkelstein *et al.*, 2002] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. *ACM Transactions on Information Systems (TOIS)*, 20(1):116–131, 2002.

[Grangier and Bengio, 2008] David Grangier and Samy Bengio. A discriminative kernel-based model to rank images from text queries. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2008.

[Haveliwala, 2003] Taher H. Haveliwala. Topic-sensitive PageRank: A context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data Engineering*, 15:784–796, 2003.

[Liben-Nowell and Kleinberg, 2003] David Liben-Nowell and Jon Kleinberg. The link prediction problem for social networks. In *Proceedings of CIKM 2003 (12th ACM International Conference on Information and Knowledge Management)*, pages 556–559, New Orleans, LA, 2003.

[Metaweb Technologies, 2010] Metaweb Technologies. Freebase Wikipedia Extraction (WEX). http://download.freebase.com/wex/, 2010.

[Miller and Charles, 1991] George A. Miller and Walter G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991.

[Navigli and Lapata, 2010] Roberto Navigli and Mirella Lapata. An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4):678–692, 2010.

[Rubenstein and Goodenough, 1965] Herbert Rubenstein and John B. Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965.

[Sarkar and Moore, 2007] Purnamrita Sarkar and Andrew Moore. A tractable approach to finding closest truncated-commute-time neighbors in large graphs. In *Proceedings of UAI 2007 (23rd Conference on Uncertainty in Artificial Intelligence)*, pages 335–343, Vancouver, BC, 2007.

[Watts and Strogatz, 1998] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[Weston *et al.*, 2011] Jason Weston, Samy Bengio, and Nicolas Usunier. WSABIE: Scaling Up to Large Vocabulary Image Annotation. In *Proceedings of IJCAI 2011 (22nd International Joint Conference on Artificial Intelligence)*, pages 2764–2770, 2011.

[Yazdani and Popescu-Belis, 2010] Majid Yazdani and Andrei Popescu-Belis. A random walk framework to compute textual semantic similarity: a unified model for three benchmark tasks. In *Proceedings of IEEE ICSC 2010 (4th IEEE International Conference on Semantic Computing)*, pages 424–429, Pittsburgh, PA, 2010.

[Yazdani and Popescu-Belis, 2013] Majid Yazdani and Andrei Popescu-Belis. Computing text semantic relatedness using the contents and links of a hypertext encyclopedia. *Artificial Intelligence*, 194:176–202, 2013.