# Top-Rank-Focused Adaptive Vote Collection for the Evaluation of Domain-Specific Semantic Models

**Pierangelo Lombardo**[1], **Alessio Boiardi**[1], **Luca Colombo**[1],
**Angelo Schiavone**[1], **and Nicolò Tamagnone**[1]
[1] Intervieweb S.r.l. (Zucchetti group), Turin, Italy
{pierangelo.lombardo, alessio.boiardi, luca.colombo,
angelo.schiavone, nicolo.tamagnone}@intervieweb.it

## Abstract

The growth of domain-specific applications of semantic models, boosted by the recent achievements of unsupervised embedding learning algorithms, demands domain-specific evaluation datasets. In many cases, content-based recommenders being a prime example, these models are required to rank words or texts according to their semantic relatedness to a given concept, with particular focus on top ranks. In this work, we give a threefold contribution to address these requirements: (*i*) we define a protocol for the construction, based on adaptive pairwise comparisons, of a relatedness-based evaluation dataset tailored on the available resources and optimized to be particularly accurate in top-rank evaluation; (*ii*) we define appropriate metrics, extensions of well-known ranking correlation coefficients, to evaluate a semantic model via the aforementioned dataset by taking into account the greater significance of top ranks. Finally, (*iii*) we define a stochastic transitivity model to simulate semantic-driven pairwise comparisons, which confirms the effectiveness of the proposed dataset construction protocol.

## 1 Introduction

In recent years, we have been witnessing a growth of Natural Language Processing (NLP) applications in a wide range of specific domains, such as recruiting (INDA; Qin et al., 2018), law (Sugathadasa et al., 2017), oil and gas (Nooralahzadeh et al., 2018), social media analysis (ALRashdi and O'Keefe, 2019), online education (Dessì et al., 2019), and biomedical (Patel et al., 2020). Embedding-based models have been playing a crucial role in this specialization, as they allow the application of the same learning algorithm to a variety of different corpora of unlabeled texts, obtaining domain-specific models (Bengio et al., 2003; Bojanowski et al., 2017; Devlin et al., 2018; Mikolov et al., 2013a, 2017, 2013b; Pennington et al., 2014).

The evaluation and validation of a domain-specialized model requires manually-annotated domain-specific datasets (Bakarov, 2018; Lastra-Díaz et al., 2019; Taieb et al., 2019). However, the construction of such datasets is a very resource-consuming process, and particular care is needed to ensure their ability to evaluate the desired features (Bakarov, 2018; Taieb et al., 2019; Wang et al., 2019). In particular, it is fundamental to carefully consider the so-called *downstream task* (i.e., the final purpose of the model), because the appropriate evaluation metric depends on this task (Bakarov, 2018; Blanco et al., 2013; Halpin et al., 2010; Rogers et al., 2018; Wang et al., 2019).

Semantic similarity and relatedness are related but distinct notions in linguistics, the first being associated with concepts which share taxonomic properties and being maximized by synonyms; on the other hand, semantically related concepts can share any kind of semantic relation, including antonym (Cai et al., 2010; Harispe et al., 2015). These notions underlie the downstream tasks of countless NLP applications, including information retrieval (Akmal et al., 2014; Chen et al., 2017; Gurevych et al., 2007; Hliaoutakis et al., 2006; Ji et al., 2017; Lopez-Gazpio et al., 2017; Srihari et al., 2000; Uddin et al., 2013), content-based recommendation (De Gemmis et al., 2008, 2015; Lops et al., 2011), semantic matching (Giunchiglia et al., 2004; Li and Xu, 2014; Wan et al., 2016), ontology learning and knowledge management (Aouicha et al., 2016a; Georgiev and Georgiev, 2018; Jiang et al., 2014; Sánchez and Moreno, 2008), and word sense disambiguation (Aouicha et al., 2016b; Patwardhan et al., 2003).

In view of the widespread of these applications, we propose a methodology to construct appropriate domain-specific datasets and metrics to assess the accuracy of relatedness and similarity estimations.



*Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 3081–3093,
November 16–20, 2020. 

In particular, due to its suitability for non-expert human annotation, we mainly focus on semantic relatedness; however, the proposed protocol can be easily extended to semantic similarity.

A standard approach to evaluate a relatedness-based model is the comparison of the semantic ranking it produces with the corresponding ranking determined from human annotations. However, the relevance of rank mismatches may depend on the involved positions; in particular, top ranks are considered more important in many contexts, two prominent examples being content-based recommenders (De Gemmis et al., 2008, 2015; Lops et al., 2011; Mladenic, 1999) and semantic matching (Giunchiglia et al., 2004; Li and Xu, 2014; Wan et al., 2016). The greater significance of top ranks compared with low ranks is actually a pretty common phenomenon, as it can be argued from the attempts to overweight the former in the context of ranking correlation (Blest, 2000; Pinto da Costa and Soares, 2005; Dancelli et al., 2013; Iman and Conover, 1987; Maturi and Abdelfattah, 2008; Shieh, 1998; Vigna, 2015; Webber et al., 2010).

Our contribution is framed within the requirement to create domain-specific datasets to evaluate semantic relatedness measure with particular focus on top ranks and is threefold. (*i*) In Section 2, we define a protocol for the construction, based on adaptive pairwise comparisons, of a relatedness-based evaluation dataset tailored on the available resources and optimized to be particularly accurate in top-rank evaluation. (*ii*) In Section 3, we define appropriate metrics to evaluate a semantic model via the aforementioned dataset by taking into account the greater significance of top ranks; the proposed metrics are extensions of well-known ranking correlation measures and they can be used to compare rankings, independently from their origin, whenever top ranks are particularly important. Finally, (*iii*) in Section 4.1, we define a stochastic model to simulate semantic-driven pairwise comparisons, whose predictions (described in Section 4.2) confirm the effectiveness of the proposed dataset construction protocol; more in detail, we adapt a stochastic transitivity model, originally defined in the context of comparative judgment, in order to make it suitable for either similarity-driven or relatedness-driven comparisons.

## 2 Dataset Construction

In this section, we describe and justify a methodology to construct a dataset for the evaluation of a domain-specific relatedness-based model. Relatedness-based evaluation – known as *intrinsic evaluation* in the context of embedding-based models – requires the construction of a dataset of human annotations, which may be collected via two different approaches. The former relies on a small group of linguistic experts to create a gold standard dataset, which is reliable but very expensive and, due to the subjectivity of relatedness and to the limited number of annotations, highly susceptible to bias and lack of statistical significance (Blanco et al., 2013; Faruqui et al., 2016). The latter relies on a large group of non-experts, typically associated with a crowdsourcing service (e.g., Amazon MTurk, ProlificAcademic, SocialSci, CrowdFlower, ClickWorker, CrowdSource), it is typically more affordable, and it has been proven to be repeatable and reliable (Blanco et al., 2013).

In the next sections we describe and justify a protocol to construct a dataset based on semantic relatedness between pairs of tokens[1] collected via a crowdsourcing approach. To simplify the reading of the paper, Figure 1 shows the main steps for the practical construction of a dataset within the proposed approach, while Table 1 reports a summary of the most frequently used symbols.

### 2.1 Token Choice

The first step in the dataset construction is the choice of the tokens among which we want to estimate the semantic relatedness. These tokens must be carefully chosen to represent the semantic areas typically involved in the downstream tasks (Bakarov, 2018; Schnabel et al., 2015). Moreover, it is well-known that models based on high-dimensional embeddings tend to incorrectly identify as a semantic nearest neighbor to almost any concept one of a few common tokens called *hubs* (Dinu et al., 2014; Feldbauer et al., 2018; Francois et al., 2007; Radovanović et al., 2010a,b). In order to detect this undesirable feature, which goes under the name of *hubness problem*, an evaluation dataset must contain a relevant amount of rare[2] tokens

[1] Hereafter, we refer as *token* to a word or a sequence of words that should be considered together, as they identify a single concept (e.g., *machine learning*).

[2] A token can be considered *rare* within a particular domain if its frequency in a corpus of domain-specific texts is, e.g., lower than 10% of the average token frequency in the corpus.

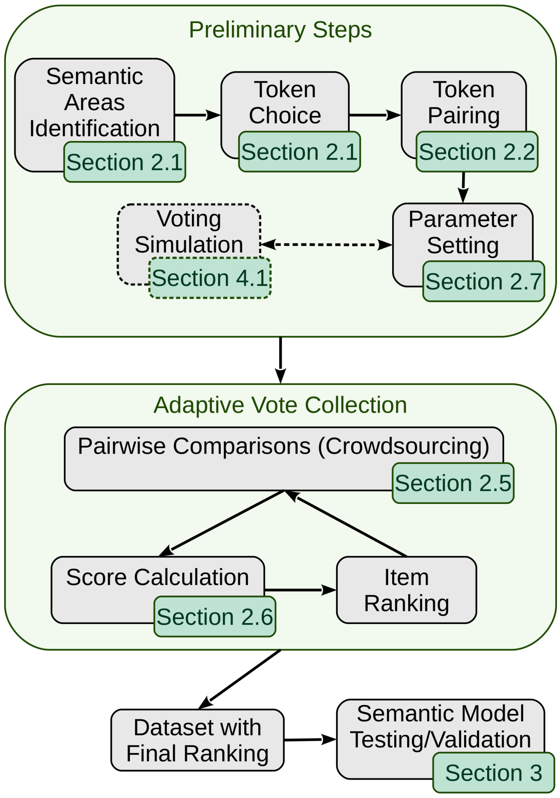


Figure 1: Main steps for the practical construction of a dataset within the proposed approach.

(Bakarov, 2018; Blanco et al., 2013). Henceforth, we consider as a concrete example a content-based recommender system in the recruitment domain (INDA); in this case, the designated tokens can be chosen among hard/soft skills, job titles, and other tokens found in resumes and job descriptions, including a relevant fraction of rare tokens.

Another potential issue of using relatedness to evaluate semantic models, is associated to lexical ambiguity, i.e., to the lack of one-to-one correspondence between tokens and meanings (Bakarov, 2018; Faruqui et al., 2016; Wang et al., 2019). To mitigate this problem, we suggest identifying a number of relevant semantic areas within the domain of interest and subdivide the tokens accordingly. For instance, *Sales & Marketing*, *Computer-related*, *Workforce*, and *Work & Welfare* are examples of semantic areas within the recruiting domain.

### 2.2 Token Pairing

The random sampling of pairs in the whole vocabulary is known to produce a large amount of unrelated pairs (Taieb et al., 2019), in contrast with the desired focus on the most related pairs. A standard approach to overcome this problem is pair selection based on either known semantic relations

Table 1: Most commonly used symbols.

| | |
|---|---|
| **Dataset Construction** | |
| $N_{\text{tok}}$ | Number of tokens |
| $N_{\text{items}}$ | Number of items (i.e., pairs of tokens) |
| $N_{\text{voters}}$ | Number of voters |
| $N_{\text{comp}}$ | Number of pairwise comparisons |
| $n_{\text{b}}$ | Number of ballots |
| $\alpha$ | Fraction of selected items |
| $i, j$ | Indices specifying the item |
| $M$ | Number of times an item is presented to the voters in each ballot |
| $k$ | Index specifying the ballot |
| $N_{\text{items}}^{(k)}$ | Number of items in ballot $k$ |
| $x_i^{(k)}$ | Borda score (i.e., win ratio) for item $i$ in ballot $k$ |
| $y_i^{(k)}$ | Rescaled score for item $i$ in ballot $k$ |
| **Evaluation Metrics** | |
| $a_i$ ($b_i$) | Rank of item $i$ according to the first (second) ranking |
| $\rho_{\text{w}}$ | Weighted version of Spearman's $\rho$ |
| $\tau_{\text{w}}$ | Weighted version of Kendall's $\tau$ |
| $w_i^a$ ($w_i^b$) | Weight associated to $a_i$ ($b_i$) |
| $n_0$ | Offset in weight calculation |
| **Stochastic Transitivity Model** | |
| $z_i$ | Underlying similarity of item $i$ |
| $v$ | Index specifying the voter |
| $o_i^{(v)}$ | Opinion of voter $v$ about item $i$ |
| $\eta_i^{(v)}$ | Gaussian-distributed random variable characterizing voter $v$ |
| $\sigma_v^*$ | Nonconformity level of voter $v$ |
| $\epsilon_v$ | Probability of oversight for voter $v$ |

or the frequency of tokens' co-occurrence within a corpus of texts. While the former information may be *a priori* unknown within the domain of interest, the latter may produce a bias in favor of distributional methods that compute relatedness based on similar knowledge sources (Taieb et al., 2019).

We suggest, therefore, separate token pairing in each of the semantic areas identified as described in Section 2.1: in this case, the relatedness distribution is substantially shifted towards larger values compared with random sampling in the whole vocabulary. This shift is shown in Figure 3 – based on a word embedding created with the word2vec algorithm (Mikolov et al., 2013a,b) trained on a corpus of resumes – where the relatedness distribution of pairs of distinct tokens selected within the same semantic area[3] (red plus) is compared with that of pairs randomly generated in the whole vocabulary (purple diamonds). The generation of all pairs of distinct tokens produces $N_{\text{tok}}(N_{\text{tok}}-1)/2$ pairs per semantic area, $N_{\text{tok}}$ being the number of tokens. Although the pairs can be reduced via random sampling, an accurate evaluation requires a large number of pairs.

### 2.3 Vote Collection

Once we have defined the pairs of tokens, which will be referred to as *items* hereafter, we want to rank them, with particular emphasis on top ranks, according to the opinions of a large number $N_{\text{voters}}$ of non experts. Due to the large number of items involved, the complete ranking of all items would be an unfeasible task for a human and it is convenient to reformulate it in terms of pairwise comparisons (Fürnkranz and Hüllermeier, 2010; Heckel et al., 2019, 2018; Jamieson and Nowak, 2011; Negahban et al., 2017; Park et al., 2015; Wauthier et al., 2013). Moreover, the complete exploration of the $N_{\text{items}}(N_{\text{items}}-1)/2$ pairs of distinct items would be extremely expensive in terms of votes; luckily enough, it has been proven to be non-necessary in many studies (Jamieson and Nowak, 2011; Negahban et al., 2017; Park et al., 2015; Wauthier et al., 2013).

Louviere and Woodworth (1991) (see also Kiritchenko and Mohammad (2017)) proposed a faster alternative to pairwise comparisons, known as *best-worst scaling*. In this case $n$-tuples (typically, $n = 4$), rather than pairs, are presented to the voter, who is required to identify the best and the worst items in each tuple, according to the relatedness of the corresponding tokens. This approach's drawbacks are a reduction, for $n > 3$, in the control on which pairs are actually checked and an increase in the complexity of each vote, which is particularly unwanted in crowdsourcing vote collections. For this reason, we rely on the standard pairwise comparison: we generate $N_{\text{comp}}$ pairs of items (as described in Sections 2.4 and 2.5), each one to be presented to one voter, who is requested to identify the item formed by the most similar tokens.

[3] More in detail, 990 pairs of distinct tokens (associated with 45 tokens) have been considered within the semantic area *Sales & Marketing*.

### 2.4 Uniform Item Selection

In our setting, each item $i$ is presented to the voters a total number of times $M_i$, and we define a score

$$x_i = \frac{n_i}{M_i}, \tag{1}$$

where $n_i$ represents the total number of times item $i$ was the winner[4] in the vote collection; note that $x_i$ corresponds to an empirical approximation of the average probability[5] – known as *Borda score* in the context of social choice theory – that item $i$ beats a randomly chosen item $j \neq i$, where the accuracy of the approximation increases as $M_i$ increases (Borda, 1784; Heckel et al., 2019).

In the absence of *a priori* knowledge on the expected scores, a reasonable approach for the data collection consists of presenting each item the same number $M_i$ of times to the voters. In this scenario, we randomly generate $N_{\text{comp}}$ pairs of items, with the constraint that $M_i = 2N_{\text{comp}}/N_{\text{items}}\ \forall i$. The only way to increase $M_i$ – which is a proxy of the accuracy of the $x_i$ score defined in Equation 1 – is, therefore, to increase the total number of comparisons $N_{\text{comp}}$.

### 2.5 Adaptive Item Selection

Crompvoets et al. (2019), Heckel et al. (2019), Heckel et al. (2018), Jamieson and Nowak (2011), and Negahban et al. (2017) proposed so-called *adaptive approaches* to increase of the efficiency of pairwise comparisons by identifying, before each comparison, the optimal pair of items to be compared based on the votes already collected, on the task to be solved (typically, finding the global ranking or a ranking-induced partition), and on assumptions on the vote distribution.

The application of an adaptive approach in our context requires two additional ingredients which, to the best of our knowledge, are still missing in the literature: (*i*) in order to avoid overfitting on the opinion of the fastest voters and to allow simultaneous voting, the choice of the pairs to be presented must occur in a few events, as each of these events causes a discontinuity in the vote collection (namely, this choice requires to suspend the vote collection when the numbers of comparisons reach the desired distribution among the voters); (*ii*) the goal is a selective increase in the precision (proxied by $M_i$) of top ranks (with no need of *a*

[4] Ties can be accounted for by defining $n_i = n_i^{\text{w}} + n_i^{\text{t}}/2$, where $n_i^{\text{w}}$ ($n_i^{\text{t}}$) represents the number of wins (ties) for item $i$.

[5] The average is intended over all voters.

*priori* knowledge on the semantic relatedness of the tokens), rather than a general improvement in the global ranking. In Section 3 we define an appropriate metric to quantify top-rank accuracy.

The key idea is to subdivide the voting procedure in $n_\mathrm{b}$ subsequent ballots in which pairwise comparisons, based on a list of pairs determined before the beginning of the ballot, are presented to the voters. During the first ballot, the pairs are randomly drawn from all items, with the constraint that each item appears $M$ times[6], while in each subsequent ballot $k$, the pairs are drawn from the $N^{(k)}_{\mathrm{items}}$ top-rank items selected according to the results of the previous ballots, with analogous constraint. More in detail, we define

$$N^{(k)}_{\mathrm{items}} = \mathrm{round}(\alpha N^{(k-1)}_{\mathrm{items}}) \sim \alpha^{k-1} N_{\mathrm{items}}, \quad (2)$$

where $\alpha$ represents the fraction of items selected at each ballot. Since each item contained in ballot $k$ appears $M$ times within the pairs of such ballot, the total number of comparisons can be written as

$$N_{\mathrm{comp}} = \sum_{k=1}^{n_\mathrm{b}} \frac{M N^{(k)}_{\mathrm{items}}}{2} \sim \frac{1-\alpha^{n_\mathrm{b}}}{2(1-\alpha)} M N_{\mathrm{items}}. \quad (3)$$

Thus, each item $i$ which survives up to the last ballot, is presented to the voters a number of times

$$M_{\mathrm{top}} = n_\mathrm{b} M \sim (1-\alpha) n_\mathrm{b} M_{\mathrm{unif}}, \quad (4)$$

where $M_{\mathrm{unif}} = 2N_{\mathrm{comp}}/N_{\mathrm{items}}$ is the number of comparisons per item in the case of uniform item selection with the same total number of comparisons; the last approximation holds whenever $\alpha^{n_\mathrm{b}} \ll 1$, i.e., when the fraction of items which survive up to the last ballot is small. According to Equation 4, the score precision for top-rank items can be increased by decreasing the fraction $\alpha$ of selected items or by increasing the number $n_\mathrm{b}$ of ballots; in Section 2.7 we discuss bounds on these values.

### 2.6 Score Calculation

At each ballot $k$ and for each item $i$ contained in $k$, we can evaluate a score $x^{(k)}_i$ defined as in Equation 1. Nonetheless, since the pool of competing items is narrowed around top ranks at each ballot, the winning chances of a given item $i$ decrease accordingly. For this reason, the expected value of $x^{(k)}_i$ is smaller than the expected value of $x^{(k-1)}_i$, and this discrepancy must be taken into account in order to average scores from different ballots.

We define therefore a rescaled score $y^{(k)}_i = f^{(k)}_{\mathrm{resc}}(x^{(k)}_i)$, where $f_{\mathrm{resc}}$ is the identity function for $k=1$, while it is obtained as a linear interpolation between $\{x^{(k)}_j\}$ and $\{\bar{y}^{(k-1)}_j\}$ for $k>1$, where $\bar{y}^{(k)}_i$ is defined as the average of all rescaled scores up to ballot $k$, i.e.,

$$\bar{y}^{(k)}_i = \frac{1}{k}\sum_{k'=1}^{k} y^{(k')}_i. \quad (5)$$

We enforce the $f^{(k)}_{\mathrm{resc}}(1) = 1$ constraint in the linear interpolation, obtaining[7], for $k>1$,

$$\begin{cases} y^{(k)}_i &= 1 - \hat{b}^{(k)} + \hat{b}^{(k)} x^{(k)}_i \\ \hat{b}^{(k)} &= \frac{\sum_j [1-x^{(k)}_j][1-\bar{y}^{(k-1)}_j]}{\sum_j [1-x^{(k)}_j]^2} \end{cases}. \quad (6)$$

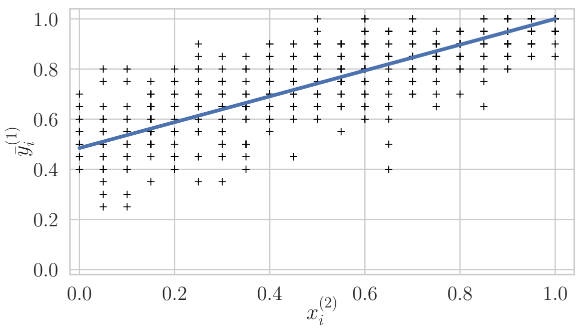


Figure 2: The coordinates of this scatterplot represent the scores $x^{(2)}_i$ and $\bar{y}^{(1)}_i = x^{(1)}_i$, for each item $i$ present in the second ballot; the data are obtained with the model described in Section 4.1 with exponential underlying similarity, $N_{\mathrm{items}} = 990$, $\alpha = 0.5$, $M = 20$, $N_{\mathrm{voters}} = 100$, $\sigma^*_v = 0.1$, $\epsilon_v = 0.01$. The blue line represents the interpolation described in the text.

Figure 2 shows an example of the interpolation on data simulated via the stochastic model described in Section 4.1. Note that Equation 5 provides a sequence of approximations of the Borda scores with accuracy increasing with $k$; top ranks are expected to survive up to the last ballot, and therefore to be highly accurate.

### 2.7 Choice of Parameter Values

We provide here heuristics to identify ranges of values for the parameters of the adaptive approach.

[6] We recommend an even value for $M$; otherwise, if both $M$ and $N_{\mathrm{items}}$ are odd, one item should appear $M+1$ times.

[7] Since the winning chances of a given item $j$ decrease at each ballot, on average, $x^{(k)}_j \le \bar{y}^{(k-1)}_j$. Heuristically, we expect that, on average, $[1-x^{(k)}_j][1-\bar{y}^{(k-1)}_j] \le [1-x^{(k)}_j]^2$, which implies $\hat{b}^{(k)} \le 1$, which in turns ensures $y^{(k)}_i \le 1$.

**Number $n_{\rm b}$ of Ballots.** We need $n_{\rm b} \geq 2$ for the adaptive approach to be meaningful, while, to limit the discontinuities in the vote collection, a reasonable upper bound is $n_{\rm b} \lesssim 10$.

**Fraction $\alpha$ of Selected Items.** As the purpose of the adaptive approach is to focus votes on top rank items, a reasonable request is to have no more than 10% of items surviving up to the last ballot, which gives an upper bound $\alpha \lesssim (0.1)^{1/(n_{\rm b}-1)}$. On the other hand, at least two items must be present in the last ballot; according to Equation 2, this implies a lower bound $\alpha \gtrsim (2/N_{\rm items})^{1/(n_{\rm b}-1)}$.

**Number $N_{\rm comp}$ of Comparisons.** To achieve the desired precision (namely, the statistical significance of the averages) on top-rank scores, a reasonable request is $M_{\rm top} \gtrsim 100$; using Equation 2 in the $\alpha^{n_{\rm b}} \ll 1$ limit, this implies $N_{\rm comp} \gtrsim 50 N_{\rm items}/[(1-\alpha)n_{\rm b}]$. On the other hand, the only upper bound on $N_{\rm comp}$ is the cost of the voting, which can be estimated from Equation 10.

**Number $M$ of Comparisons per Ballot.** $M$ can be computed from Equation 3 once the other parameters have been fixed. Note the existence of two competing phenomena: (*i*) decreasing $M$ (i.e., increasing $\alpha$) we increase the fluctuations in the $x_i$ scores defined in Equation 1; (*ii*) for given $x_i$ fluctuations, decreasing $\alpha$ we increase the probability of top ranks' premature loss due to stricter selection. A rigorous derivation of the optimal values of $M$ and $\alpha$ as a trade-off between these phenomena is beyond the scope of this section, as the bounds discussed above provide heuristic ranges.

## 3 Evaluation Metrics

Although the approximations of the Borda scores described in Sections 2.4 and 2.5 can be thought as estimates of the semantic relatedness, we rely on rankings rather than scores to avoid inconsistency issues that frequently emerge in score comparisons (Ammar and Shah, 2011; Negahban et al., 2017).

Kendall (1948) proposed the quite general form for a ranking correlation coefficient

$$\Gamma = \frac{\sum_{i,j} a_{ij} b_{ij}}{\sqrt{\sum_{ij} a_{ij}^2 \sum_{ij} b_{ij}^2}}, \tag{7}$$

where $a_{ij}$ (resp., $b_{ij}$) is a matrix that depends on the first (second) ranking to be compared, with indices $i, j$ running over all items. This definition contains Spearman's $\rho$ (Spearman, 1961) as a particular case with $a_{ij} = a_j - a_i$ and $b_{ij} = b_j - b_i$, while Kendall's $\tau$ (Kendall, 1938; Kruskal, 1958) is obtained with $a_{ij} = \mathrm{sign}(a_j - a_i)$ and $b_{ij} = \mathrm{sign}(b_j - b_i)$, where $\{a_i\}$ and $\{b_i\}$ are the rankings to be compared.

In order to take into account the larger importance of top ranks in our context, we define weighted versions of $\rho$ and $\tau$, with increasing weight at the increasing of the rank position. Namely, we define (*i*) $a_{ij} = \sqrt{w_i w_j}\,(a_j - a_i)$, $b_{ij} = \sqrt{w_i w_j}\,(b_j - b_i)$ and (*ii*) $a_{ij} = \sqrt{w_i w_j}\,\mathrm{sign}(a_j - a_i)$, $b_{ij} = \sqrt{w_i w_j}\,\mathrm{sign}(b_j - b_i)$, where $w_i$ is the normalized weight associated to the $i$-th position in the rankings. These coefficients can be rewritten respectively as

$$\begin{aligned} \rho_{\rm w} &= \frac{\sum_i w_i (a_i - \bar{a})(b_i - \bar{b})}{\sigma_a \sigma_b} \\ \tau_{\rm w} &= \frac{\sum_{ij} w_i w_j\, \mathrm{sign}(a_j - a_i)\, \mathrm{sign}(b_j - b_i)}{Z(\{w_i\})} \end{aligned}, \tag{8}$$

where $\bar{a} = \sum_i w_i a_i$, $\sigma_a^2 = \sum_i w_i (a_i^2 - \bar{a}^2)$, $\bar{b} = \sum_i w_i b_i$, $\sigma_b^2 = \sum_i w_i (b_i^2 - \bar{b}^2)$, while $Z(\{w_i\})$ is a normalization factor, which corresponds to $1 - \sum_i w_i^2$ in the absence of ties; these metrics have been emerging, albeit some notation differences, as extensions of the $\rho$ and $\tau$ coefficients to take into account the larger importance of top ranks (Pinto da Costa and Soares, 2005; Dancelli et al., 2013; Vigna, 2015).

Different weighting schemes have been proposed in the literature (Dancelli et al., 2013; Vigna, 2015); here we adopt the additive scheme

$$w_i = \frac{w_i^a + w_i^b}{\sum_j (w_j^a + w_j^b)}, \tag{9}$$

with $w_i^a = f(a_i)$ and $w_i^b = f(b_i)$, where $f(n)$ is a monotonically decreasing function, in view of its ability in discriminating different rankings even when they only differ by the exchange of a top rank and a low rank (Dancelli et al., 2013).

A common choice is $f(n) = 1/n$ (Dancelli et al., 2013; Vigna, 2015); however, in the large $N_{\rm items}$ limit, it causes the divergence of the denominator in Equation 9 and makes thus any $w_i$ negligible. This phenomenon is responsible for the decreased sensitivity on top ranks, observed by Dancelli et al. (2013), in case of long rankings. For this reason, we prefer to use $f(n; n_0) = 1/(n + n_0)^2$, where the offset $n_0$ has been introduced to control the weight fraction associated to the first rank in the large $N_{\rm items}$ limit, i.e., $R(n_0) = f(1; n_0)/\sum_{n=1}^{\infty} f(n; n_0)$, which can be expressed as $R(n_0) = 1/[(n_0 + 1)^2\, \psi^{(1)}(n_0 + 1)]$, where $\psi^{(1)}(x)$ is the first derivative of the digamma function. With this choice, both $\rho_\mathrm{w}$ and $\tau_\mathrm{w}$ defined in Equation 8 represent a family of correlation coefficients, depending on the value of $n_0$, whose choice depends on the particular task (namely, on the relative importance of the first rank). The value $n_0 = 0$ causes an extremely high sensitivity on the first rank ($R(0) \sim 0.61$), which may be excessive; hereafter, we focus therefore on the value $n_0 = 2$, which appears to be a reasonable trade-off ($R(2) \sim 0.28$) that allows focusing on the first rank while avoiding neglecting other ranks.

The metrics $\rho_\mathrm{w}$ and $\tau_\mathrm{w}$ are suitable to compare rankings, whenever top ranks are particularly important; in particular, they can be used to evaluate a semantic model using a dataset produced as described in this paper.

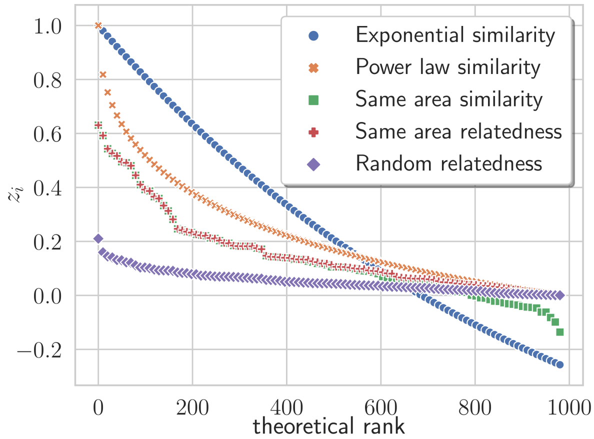


Figure 3: We represent the three underlying similarity distributions described in Section 4.1.1 and the two relatedness distributions described in Section 2.2; relatedness is quantified by $|\cos\theta|$, where $\theta$ is the angle between the corresponding vectors in the embedding.

## 4 Evaluation of the Data-Collection Framework

The collection of human annotations to construct a domain-specific dataset is a resource-consuming process, even within the proposed optimized data collection approach, whose *person-hours* cost can be estimated as

$$C \sim \bar{t}_\mathrm{comp} N_\mathrm{comp}, \tag{10}$$

where $\bar{t}_\mathrm{comp}$ is the average time needed for a single comparison. For this reason, in Section 4.1, we define a stochastic model for semantic pairwise comparisons, which can be used to simulate the voting before the collection of human annotations, e.g., for checking or tuning the parameters of the data collection approach. This stochastic model will be used in Section 4.2 to compare the effectiveness of the adaptive and the uniform approaches, using the metrics defined in Section 3.

### 4.1 Semantic Pairwise Comparisons

We want to model $N_\mathrm{voters}$ voters to whom are proposed $N_\mathrm{comp}$ pairwise comparisons and who are asked to identify the item containing the most semantically related tokens. The model will be used to reconstruct an approximate ranking of the items.

For the sake of mathematical simplicity, we firstly focus on similarity-driven comparisons, where the similarity $z$ takes value in the symmetric interval $[-1, 1]$, where $z = 1$, $0$, and $-1$ correspond respectively to synonyms, unrelated tokens, and antonyms. The model will eventually be adapted to semantic relatedness by using the fact that, since antonyms correspond to semantically related tokens (Cai et al., 2010; Harispe et al., 2015), the absolute value $|z|$ is a reasonable proxy for semantic relatedness.

#### 4.1.1 Similarity-Driven Comparisons.

A convenient way to model similarity-driven pairwise comparisons assumes the existence of an underlying (unknown) similarity distribution $\{z_i\}$, which determines the theoretical ranks of the items, which in turn can be compared with the ranks estimated via the model. We consider here three examples:(*i*) an exponential $z_i = 2\exp(-i/N_\mathrm{items}) - 1$, (*ii*) a power law $z_i = 2/(1 + \sqrt{i/N_\mathrm{items}}) - 1$, and (*iii*) the distribution of the cosine similarity[8] between pairs of tokens in the word embedding described in Section 2.2; these distributions are represented in Figure 3.

A fundamental aspect to be considered in modelling similarity-driven pairwise comparisons is the task's subjectivity, as many potential linguistic, psychological, and social factors could introduce biases (Bakarov, 2018; Faruqui et al., 2016; Gladkova and Drozd, 2016). A possible approach to account for this problem is via a stochastic transitivity model, firstly introduced in the context of comparative judgment of physical stimuli by Thurstone (1927) (see also Cattelan, 2012; Ennis, 2016); this model describes the opinion $o_i^{(v)}$ of voter $v$ about item $i$ as a stochastic function

[8] Despite a certain ambiguity observed by Faruqui et al. (2016), cosine similarity is typically considered a proxy of semantic similarity (Auguste et al., 2017; Banjade et al., 2015).

$o_i^{(v)} = z_i + \sigma_v \eta_i^{(v)}$, where $z_i$ is the underlying similarity, $\eta_i^{(v)}$ is a Gaussian-distributed random variable with zero mean and unit variance, while $\sigma_v$ represents the *nonconformity amplitude*, i.e., the discrepancy between the voter's opinion and the underlying similarity.[9]

Here we define a modified version of the Thurstonian model, with a stochastic amplitude $\sigma_v$ depending on the underlying similarity, so that

$$o_i^{(v)} = F(z_i + \sigma_v(z_i)\,\eta_i^{(v)}), \tag{11}$$

where $F(x) = \max(-1, \min(1, x))$ has been introduced to enforce the constraint $-1 \le o_i^{(v)} \le 1$, analogous to the one discussed above for $z$. In the absence of $z_i$ dependence in the nonconformity amplitude, the probability $P_{\text{out}}(z_i)$ to have $z_i + \sigma_v \eta_i$ outside the interval $[-1, 1]$ would tend to $1/2$ as $z_i$ approaches one of the boundaries of the interval, causing, due to the $F$ constraint, the collapse of a relevant fraction of opinion $o_i$ to either $-1$ or 1. This degeneracy can be avoided with $\sigma_v(z_i)$ proportional to $1 - z_i$ and $1 + z_i$ as $z_i$ approaches 1 and $-1$ respectively. Here we consider the simplest form with these features, i.e., $\sigma_v(z_i) = \sigma_v^* (1 - z_i^2)$, which makes particularly sense in our context, where each item $i$ represents a pair of tokens, and the closest the similarity is to $z_i = 1$ ($z_i = -1$), the higher is the relation (the opposition) between the tokens in the corresponding pair, and the stronger is expected to be the agreement in the voters' opinions on their similarity.

In order to increase the accuracy of the model, we introduce another source of randomness that represents the *distraction level* of the voter, i.e., its tendency – observed, e.g., by Bakarov (2018) and Bruni et al. (2014) – to unintentionally vote for the item perceived as lower rank. This tendency is accounted for by assuming that the result of a pairwise comparison presented to voter $v$ is actually the item with the highest perceived score $o_i^{(v)}$ with probability $1 - \epsilon_v$ (with $\epsilon_v \ll 1$), while the other item is voted (oversight) with probability $\epsilon_v$.

The proposed model depends on the underlying similarity distribution[10], on the number of voters, on the random variables $\eta_i^{(v)}$, and on the voter-distinctive parameters $\sigma_v^*$ and $\epsilon_v$, whose distribution could be experimentally determined by analyzing human voting. In the absence of such analysis, it seems reasonable to uniformly draw $\sigma_v^*$ and $\epsilon_v$ from ranges covering one order of magnitude to encompass human variability; heuristic upper bounds are $\epsilon_v \lesssim 0.05$ and $\sigma_v^* \lesssim 0.2$, as oversights are supposed to be rare, and the probability $P_{\text{out}}(0)$ that two completely unrelated tokens ($z_i = 0$) are deliberately considered as maximally related ($o_i^{(v)} = \pm 1$) should be extremely low: the aforementioned bound corresponds indeed to $P_{\text{out}}(0) \lesssim 0.001\%$.

#### 4.1.2 Relatedness-Driven Comparisons.

As discussed in Section 4.1, we consider the absolute value of similarity as a proxy of relatedness. The model defined in Section 4.1.1 is thus extended to relatedness-driven comparisons by (*i*) including an absolute value in Equation 11, so that $o_i^{(v)} = |F(z_i + \sigma_v^*(z_i)\,\eta_i^{(v)})|$ and (*ii*) defining the theoretical rank of item $i$ according to $|z_i|$.

### 4.2 Results

We estimated the accuracy of a data collection approach by comparing, via the metrics defined in Section 3, the ranking that it produces with the underlying theoretical ranks. We considered the semantic area described in Section 2.2, containing 990 items, and we simulated a relatedness-driven data collection based on (*i*) the adaptive approach described in Section 2.5, with $N_{\text{comp}} = 39000$, $M = 20$, $\alpha = 0.5$, $n_{\text{b}} = 7$ and (*ii*) the uniform approach described in Section 2.4, with the same total number of comparisons. The voting was simulated with the stochastic model described in Section 4.1, with $N_{\text{voters}} = 100$ and based on all three discussed distributions for the underlying similarity; for each voter $v$, the nonconformity level $\sigma_v^*$ and the distraction level $\epsilon_v$ were randomly chosen in the intervals $[0.02, 0.2]$ and $[0.005, 0.05]$ respectively, while each $\eta_i^{(v)}$ was randomly drawn from a normal distribution with zero mean and unit variance. Each simulation was repeated 50 times (by resampling all voters' parameters $\sigma_v^*$, $\epsilon_v$, and $\eta_i^{(v)}$ at each simulation) in order to obtain statistically significant results.

The code for our experiments is available at https://github.com/interviewweb-datascience/adaptive-comp and was run on a local machine equipped with an Intel Core i7-7700HQ (2.80GHz x8), with

[9] Contrary to the original formulation, no covariance terms are present here, as the voters are supposed to be non-interacting. Moreover, in the original formulation, voters and items are respectively referred to as *judges* and *stimuli*.

[10] However, as shown in Table 2, the dependence is mild.

Table 2: Mean ± standard deviation (unbiased estimation over 50 simulations of relatedness-driven comparisons, as described in text) for $\rho_w$ and $\tau_w$ metrics, defined in Equation 8, and for Spearman's $\rho$ and Kendall's $\tau$ coefficients.

| | | $\rho_w$ | $\tau_w$ | $\rho$ | $\tau$ |
|---|---|---|---|---|---|
| Exponential | Uniform | $.778 \pm .058$ | $-0.11 \pm .20$ | $.8097 \pm .0088$ | $.6265 \pm .0091$ |
| | Adaptive | $.9452 \pm .0028$ | $.66 \pm .17$ | $.8015 \pm .0087$ | $.6330 \pm .0098$ |
| Power Law | Uniform | $.800 \pm .062$ | $-0.11 \pm .20$ | $.9713 \pm .0013$ | $.8491 \pm .0035$ |
| | Adaptive | $.9800 \pm .0014$ | $.63 \pm .18$ | $.9632 \pm .0019$ | $.8406 \pm .0040$ |
| Embedding | Uniform | $.741 \pm .058$ | $-0.11 \pm .21$ | $.7229 \pm .0078$ | $.5463 \pm .0072$ |
| | Adaptive | $.9146 \pm .0042$ | $.73 \pm .12$ | $.7258 \pm .0093$ | $.5611 \pm .0097$ |

average runtimes of 30.9 s and 33.2 s respectively for the adaptive and the uniform approaches. The results of the simulations are presented in Table 2, which contains, as measures of the accuracy of the proposed approaches, the $\rho_w$ and $\tau_w$ coefficients defined in Equation 8 and discussed in Section 3; in order to check the overall rank accuracy, we also report the standard Spearman's $\rho$ and Kendall's $\tau$ coefficients. For each coefficient, we report the average value and the unbiased estimator of the standard deviation over the 50 simulations. The adaptive approach, compared with the uniform approach, determines a relevant increase in both $\rho_w$ and $\tau_w$ for any of the underlying similarity distributions considered, with no relevant changes in the overall rank precision measured by $\rho$ and $\tau$. Moreover, the results suggest that the proposed stochastic model is robust for changes in the underlying similarity distribution.

Figure 4 displays the scores $x_i^{(k)}$ calculated in the first 5 ballots and the final approximation $\bar{y}_i$, obtained in a simulation based on the adaptive approach with exponential underlying similarity and the parameters described above; the figure clearly shows that, as desired, the $\bar{y}_i$ precision is substantially larger for top ranks.

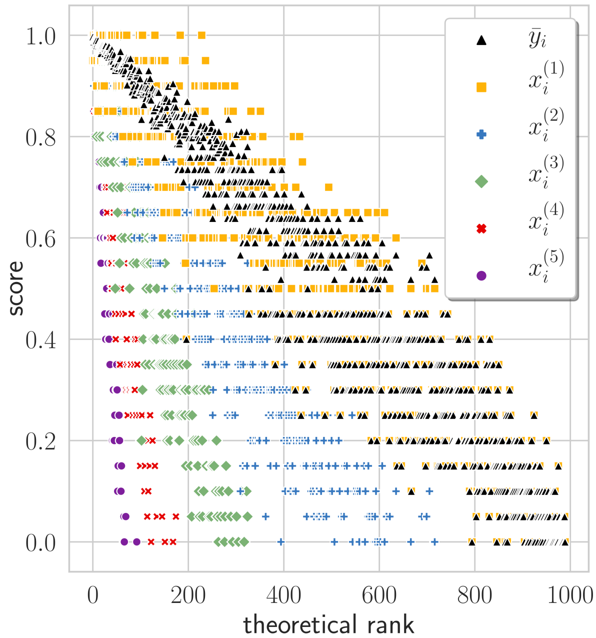


Figure 4: The scores $x_i^{(k)}$ calculated in the first 5 ballots and the final approximation $\bar{y}_i$ are displayed as functions of the theoretical ranks. All values are obtained in the simulation described in the text.

## 5 Conclusion & Future Work

In this paper, we provided a protocol for the construction – based on adaptive pairwise comparisons and tailored on the available resources – of a dataset, which can be used to test or validate any relatedness-based domain-specific semantic model and which is optimized to be particularly accurate in top-rank evaluation. Moreover, we defined the metrics $\rho_w$ and $\tau_w$, extensions of well-known ranking correlation coefficients, to evaluate a semantic model via the aforementioned dataset by taking into account the greater significance of top ranks. Finally, we defined a stochastic transitivity model to simulate semantic-driven pairwise comparisons, which allows tuning the parameters of the data collection approach and which confirmed a significant increase in the performance metrics $\rho_w$ and $\tau_w$ of the proposed adaptive approach compared with the uniform approach (see Table 2).

As future work, we plan to collect human annotations (*i*) to test the proposed data collection approach on real data and (*ii*) to assess the validity and estimate the parameters of the proposed stochastic transitivity model. Additional future investigations may include a deeper analysis of the mathematical and statistical properties of the weighted coefficients $\rho_w$, $\tau_w$, as well as a rigorous derivation of the optimal values for the parameters of the data collection approach.

## References


Suriati Akmal, Li-Hsing Shih, and Rafael Batres. 2014. Ontology-based similarity for product information retrieval. *Computers in Industry*, 65(1):91–107.

Reem ALRashdi and Simon O'Keefe. 2019. Deep learning and word embeddings for tweet classification for crisis response. *arXiv preprint arXiv:1903.11024*.

Ammar Ammar and Devavrat Shah. 2011. Ranking: Compare, don't score. In *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 776–783. IEEE.

Mohamed Ben Aouicha, Mohamed Ali Hadj Taieb, and Malek Ezzeddine. 2016a. Derivation of "is a" taxonomy from wikipedia category graph. *Engineering Applications of Artificial Intelligence*, 50:265–286.

Mohamed Ben Aouicha, Mohamed Ali Hadj Taieb, and Hania Ibn Marai. 2016b. Wsd-tic: word sense disambiguation using taxonomic information content. In *International Conference on Computational Collective Intelligence*, pages 131–142. Springer.

Jeremy Auguste, Arnaud Rey, and Benoit Favre. 2017. Evaluation of word embeddings against cognitive processes: primed reaction times in lexical decision and naming tasks. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 21–26.

Amir Bakarov. 2018. A survey of word embeddings evaluation methods. *arXiv preprint arXiv:1801.09536*.

Rajendra Banjade, Nabin Maharjan, Nobal B Niraula, Vasile Rus, and Dipesh Gautam. 2015. Lemon and tea are not similar: Measuring word-to-word similarity by combining different methods. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 335–346. Springer.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Roi Blanco, Harry Halpin, Daniel M Herzig, Peter Mika, Jeffrey Pound, Henry S Thompson, and Thanh Tran. 2013. Repeatable and reliable semantic search evaluation. *Journal of web semantics*, 21:14–29.

David C Blest. 2000. Theory & methods: Rank correlation—an alternative measure. *Australian & New Zealand Journal of Statistics*, 42(1):101–111.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

JC de Borda. 1784. Mémoire sur les élections au scrutin. *Histoire de l'Academie Royale des Sciences pour 1781 (Paris, 1784)*.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of artificial intelligence research*, 49:1–47.

Songmei Cai, Zhao Lu, and Junzhong Gu. 2010. An effective measure of semantic similarity. In *Advances in Wireless Networks and Information Systems*, pages 9–17. Springer.

Manuela Cattelan. 2012. Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science*, pages 412–433.

Fuzan Chen, Chenghua Lu, Harris Wu, and Minqiang Li. 2017. A semantic similarity measure integrating multiple conceptual relationships for web service discovery. *Expert Systems with Applications*, 67:19–31.

Joaquim Pinto da Costa and Carlos Soares. 2005. A weighted rank measure of correlation. *Australian & New Zealand Journal of Statistics*, 47(4):515–529.

Elise AV Crompvoets, Anton A Béguin, and Klaas Sijtsma. 2019. Adaptive pairwise comparison for educational measurement. *Journal of Educational and Behavioral Statistics*, page 1076998619890589.

Livia Dancelli, Marica Manisera, and Marika Vezzoli. 2013. On two classes of weighted rank correlation measures deriving from the spearman's $\rho$. In *Statistical Models for Data Analysis*, pages 107–114. Springer.

Marco De Gemmis, Pasquale Lops, Cataldo Musto, Fedelucio Narducci, and Giovanni Semeraro. 2015. Semantics-aware content-based recommender systems. In *Recommender Systems Handbook*, pages 119–159. Springer.

Marco De Gemmis, Pasquale Lops, Giovanni Semeraro, and Pierpaolo Basile. 2008. Integrating tags in a semantic content-based recommender. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 163–170.

Danilo Dessì, Mauro Dragoni, Gianni Fenu, Mirko Marras, and Diego Reforgiato Recupero. 2019. Evaluating neural word embeddings created from online course reviews for sentiment analysis. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pages 2124–2127.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2014. Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*.

Daniel M Ennis. 2016. *Thurstonian models: Categorical decision making in the presence of noise*. Institute for Perception.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. *arXiv preprint arXiv:1605.02276*.

Roman Feldbauer, Maximilian Leodolter, Claudia Plant, and Arthur Flexer. 2018. Fast approximate hubness reduction for large high-dimensional data. In *2018 IEEE International Conference on Big Knowledge (ICBK)*, pages 358–367. IEEE.

Damien Francois, Vincent Wertz, and Michel Verleysen. 2007. The concentration of fractional distances. *IEEE Transactions on Knowledge and Data Engineering*, 19(7):873–886.

Johannes Fürnkranz and Eyke Hüllermeier. 2010. Preference learning and ranking by pairwise comparison. In *Preference learning*, pages 65–82. Springer.

Georgi V Georgiev and Danko D Georgiev. 2018. Enhancing user creativity: Semantic measures for idea generation. *Knowledge-Based Systems*, 151:1–15.

Fausto Giunchiglia, Pavel Shvaiko, and Mikalai Yatskevich. 2004. S-match: an algorithm and an implementation of semantic matching. In *European semantic web symposium*, pages 61–75. Springer.

Anna Gladkova and Aleksandr Drozd. 2016. Intrinsic evaluations of word embeddings: What can we do better? In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 36–42.

Iryna Gurevych, Christof Müller, and Torsten Zesch. 2007. What to be?-electronic career guidance based on semantic relatedness. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 1032–1039.

Harry Halpin, Daniel M Herzig, Peter Mika, Roi Blanco, Jeffrey Pound, Henry Thompon, and Duc Thanh Tran. 2010. Evaluating ad-hoc object retrieval. In *IWEST@ ISWC*.

Sébastien Harispe, Sylvie Ranwez, and Stefan Janaqi. 2015. *Semantic similarity from natural language and ontology analysis*. Morgan & Claypool Publishers.

Reinhard Heckel, Nihar B Shah, Kannan Ramchandran, Martin J Wainwright, et al. 2019. Active ranking from pairwise comparisons and when parametric assumptions do not help. *The Annals of Statistics*, 47(6):3099–3126.

Reinhard Heckel, Max Simchowitz, Kannan Ramchandran, and Martin J Wainwright. 2018. Approximate ranking from pairwise comparisons. *arXiv preprint arXiv:1801.01253*.

Angelos Hliaoutakis, Giannis Varelas, Epimenidis Voutsakis, Euripides GM Petrakis, and Evangelos Milios. 2006. Information retrieval by semantic similarity. *International journal on semantic Web and information systems (IJSWIS)*, 2(3):55–73.

Ronald L Iman and WJ Conover. 1987. A measure of top–down correlation. *Technometrics*, 29(3):351–357.

INDA. INtelligent Data Analysis for HR (*https://inda.ai*).

Kevin G Jamieson and Robert Nowak. 2011. Active ranking using pairwise comparisons. In *Advances in Neural Information Processing Systems*, pages 2240–2248.

Xiaonan Ji, Alan Ritter, and Po-Yin Yen. 2017. Using ontology-based semantic similarity to facilitate the article screening process for systematic reviews. *Journal of biomedical informatics*, 69:33–42.

Yong Jiang, Xinmin Wang, and Hai-Tao Zheng. 2014. A semantic similarity measure based on information distance for ontology alignment. *Information Sciences*, 278:76–87.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Maurice George Kendall. 1948. Rank correlation methods.

Svetlana Kiritchenko and Saif M Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. *arXiv preprint arXiv:1712.01765*.

William H Kruskal. 1958. Ordinal measures of association. *Journal of the American Statistical Association*, 53(284):814–861.

Juan J Lastra-Díaz, Josu Goikoetxea, Mohamed Ali Hadj Taieb, Ana García-Serrano, Mohamed Ben Aouicha, and Eneko Agirre. 2019. A reproducible survey on word embeddings and ontology-based methods for word similarity: linear combinations outperform the state of the art. *Engineering Applications of Artificial Intelligence*, 85:645–665.

Hang Li and Jun Xu. 2014. Semantic matching in search. *Foundations and Trends in Information retrieval*, 7(5):343–469.

Iñigo Lopez-Gazpio, Montse Maritxalar, Aitor Gonzalez-Agirre, German Rigau, Larraitz Uria, and Eneko Agirre. 2017. Interpretable semantic textual similarity: Finding and explaining differences between sentences. *Knowledge-Based Systems*, 119:186–199.

Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. 2011. Content-based recommender systems: State of the art and trends. In *Recommender systems handbook*, pages 73–105. Springer.

Jordan J Louviere and George G Woodworth. 1991. Best-worst scaling: A model for the largest difference judgments. *University of Alberta: Working Paper*.

Tahani A Maturi and Ezz H Abdelfattah. 2008. A new weighted rank correlation. *Journal of mathematics and statistics.*, 4(4):226–230.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2017. Advances in pre-training distributed word representations. *arXiv preprint arXiv:1712.09405*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Dunja Mladenic. 1999. Text-learning and related intelligent agents: a survey. *IEEE intelligent systems and their applications*, 14(4):44–54.

Sahand Negahban, Sewoong Oh, and Devavrat Shah. 2017. Rank centrality: Ranking from pairwise comparisons. *Operations Research*, 65(1):266–287.

Farhad Nooralahzadeh, Lilja Øvrelid, and Jan Tore Lønning. 2018. Evaluation of domain-specific word embeddings using knowledge resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Dohyung Park, Joe Neeman, Jin Zhang, Sujay Sanghavi, and Inderjit Dhillon. 2015. Preference completion: Large-scale collaborative ranking from pairwise comparisons. In *International Conference on Machine Learning*, pages 1907–1916.

Rashmi Patel, Jessica Irving, Matthew Taylor, Hitesh Shetty, Megan Pritchard, Robert Stewart, Paolo Fusar-Poli, and Philip McGuire. 2020. T109. traversing the transdiagnostic gap between depression, mania and psychosis with natural language processing. *Schizophrenia Bulletin*, 46(Supplement_1):S272–S273.

Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen. 2003. Using measures of semantic relatedness for word sense disambiguation. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 241–257. Springer.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Chuan Qin, Hengshu Zhu, Tong Xu, Chen Zhu, Liang Jiang, Enhong Chen, and Hui Xiong. 2018. Enhancing person-job fit for talent recruitment: An ability-aware neural network approach. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 25–34.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010a. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(Sep):2487–2531.

Milos Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010b. On the existence of obstinate results in vector space models. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 186–193.

Anna Rogers, Shashwath Hosur Ananthakrishna, and Anna Rumshisky. 2018. What's in your embedding, and how it predicts task performance. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2690–2703.

David Sánchez and Antonio Moreno. 2008. Learning non-taxonomic relationships from web documents for domain ontology construction. *Data & Knowledge Engineering*, 64(3):600–623.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 298–307.

Grace S Shieh. 1998. A weighted kendall's tau statistic. *Statistics & probability letters*, 39(1):17–24.

Charles Spearman. 1961. The proof and measurement of association between two things.

Rohini K Srihari, Zhongfei Zhang, and Aibing Rao. 2000. Intelligent indexing and semantic retrieval of multimodal documents. *Information Retrieval*, 2(2-3):245–275.

Keet Sugathadasa, Buddhi Ayesha, Nisansa de Silva, Amal Shehan Perera, Vindula Jayawardana, Dimuthu Lakmal, and Madhavi Perera. 2017. Synergistic union of word2vec and lexicon for domain specific semantic similarity. In *2017 IEEE International Conference on Industrial and Information Systems (ICIIS)*, pages 1–6. IEEE.

Mohamed Ali Hadj Taieb, Torsten Zesch, and Mohamed Ben Aouicha. 2019. A survey of semantic relatedness evaluation datasets and procedures. *Artificial Intelligence Review*, pages 1–42.

Louis L Thurstone. 1927. A law of comparative judgment. *Psychological review*, 34(4):273.

Mohammed Nazim Uddin, Trong Hai Duong, Ngoc Thanh Nguyen, Xin-Min Qi, and Geun Sik Jo. 2013. Semantic similarity measures for enhancing information retrieval in folksonomies. *Expert Systems with Applications*, 40(5):1645–1653.

Sebastiano Vigna. 2015. A weighted correlation index for rankings with ties. In *Proceedings of the 24th international conference on World Wide Web*, pages 1166–1176.

Shengxian Wan, Yanyan Lan, Jiafeng Guo, Jun Xu, Liang Pang, and Xueqi Cheng. 2016. A deep architecture for semantic matching with multiple positional sentence representations. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Bin Wang, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C-C Jay Kuo. 2019. Evaluating word embedding models: methods and experimental results. *APSIPA Transactions on Signal and Information Processing*, 8.

Fabian Wauthier, Michael Jordan, and Nebojsa Jojic. 2013. Efficient ranking from pairwise comparisons. In *International Conference on Machine Learning*, pages 109–117.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38.