Wordnet-based Similarity Measure for Polish Short Texts

Maciej Piasecki, Anna Gut

G4.19 Research Group, Institute of Informatics
Wrocław University of Technology, Wrocław, Poland
maciej.piasecki@pwr.wroc.pl,anna.anaaa.gut@gmail.com

Abstract

We present a method for computing semantic similarity of Polish text with main focus given to short texts. We took into account the limited set of language tools for Polish, especially the not sufficient development of syntactic and semantic parsers. plWordNet is used to construct meaning representations for words in such a way that different words of the similar meaning receive similar representation. The use of a Word Sense Disambiguation tool for Polish brought positive results in one of the method variants in spite of the limited accuracy of the tool. The proposed measures have been compared with the manual evaluation of sentence pairs. The measures were also applied as a part of the Question Answering system. Improved performance of answer finding was achieved in several types of tests.

1. Introduction

Computing text-to-text similarity is a key issue for many applications. It is getting more difficult if the compared texts are short, at least one of them. A good text similarity measure should go beyond string comparison and should be based on semantic content of texts. The problem is naturally separated into two similarity levels: words and text structures. The depth of the analysis of text structures is determined by the available language tools. Methods for calculating semantic similarity of short texts can be divided (Achananuparp et al., 2008) into three main types: methods based on overlap of words, TF.IDF-based methods (Salton and McGill, 1986), and linguistic measures.

Methods of the first two groups represent document as a bag of words (collection) and ignore linguistic structures. In first group overlap of words between sentences is calculated, e.g. with the help of Jacquard or Dice measure. The comparison is done mostly on the level of text words but filtered by a stop list or limited to selected grammatical classes. In the case of Polish rich inflection and weakly constrained word order require preprocessing on the morpho-syntactic level and mapping words onto their lemmas. TF.IDF measures refer to the well known and mostly effective vector model for Information Retrieval.

Linguistic measures are a heterogeneous group and explore information provided by the available language resources and tools. Several approaches in this group are based only on word similarity, e.g. (Corley and Mihalcea, 2005), (Li et al., 2006), (Bär et al., 2012), taking as input text words. (Corley and Mihalcea, 2005) proposed a mixed measure based on calculating similarity of words between two texts with more weight given to more specific words. Any word-to-word similarity can be used. This approach was tested on word similarity measures based on Princeton WordNet (Fellbaum, 1998). IDF was used as a factor representing word specificity. This method has been further extended in (Mihalcea et al., 2006) with a larger number of word similarity measures, including measures based on text corpora, not only on WordNet.

(Li et al., 2006) considered not only the similarity of words but also the order of their occurrence. (Liu and Wang., 2014) expanded wordnet-based word similarity to computing the similarity of sentences. Word similarity measure proposed in (Pourgholamali and Kahani, 2012) was used. The semantic similarity of sentences was computed in four steps. First text words are mapped onto concepts in ontology. The identified nodes are expanded, i.e. direct descending nodes are added and all ancestors along one street path. The nodes are weighted according to the path distance from the directly mapped node. Finally the constructed vectors are compared with the cosine measure.

(Corley and Mihalcea, 2005) proposed merging together several measures. (Bär et al., 2012) tested large number of measures and selected those producing the best results. In addition to the text similarity measures, they calculated also the longest common substring and the number of n-grams based on characters and words.

Tree Edit Distance has been also applied in measuring the semantic similarity of texts, e.g. (Kouylekov and Magnini, 2005). (Punyakanok et al., 2004) achieved higher accuracy using graphs produced by the dependency parser.

Methods based on the comparison of syntactic structures can be applied to Polish to a limited extent. Dependency parsers for Polish express quite substantial error rate, other parsers do not provide disambiguation of structures or have limited coverage. WordNet is associated with a corpus including manually disambiguated word senses that allows for collecting the word sense frequencies. Such data that are a basis for many word similarity measures are not accessible for Polish. Thus we concentrated on optimal use of the available language resources, e.g. plWordNet, and tools, e.g. robust morpho-syntactic taggers.

Our goal was to develop a method for computing similarity of short texts in Polish aimed at capturing the similarity of the information conveyed by texts regardless of particular words used. We wanted to base the description of the lexical meanings on plWordNet and to apply a set of possibly simple language tools, without the need of referring to some form of parsing.

2. Wordnet-based Text Similarity Measure

The same message can be often expressed using different words. In longer texts different synonyms occur interchangeably across a single document, while in short texts one synonym is often enough. When two bag-of-word representations of synonymous short texts are compared, the

mismatch is very likely. In order to check if two short texts are about the same topic, we need to abstract from the exact words used in them.

Facing the lack of robust parsers for Polish, we selected bag of words representation. In UKP system (Bär et al., 2012) a thesaurus built by the means of crowdsourcing was utilised. In the case of Polish, plWordNet – a very large Polish wordnet (Maziarz et al., 2014) can be used instead. Texts to be compared are morpho-syntactically tagged and lemmatised. A lemma can correspond to several lexical meanings represented by plWordNet synsets (sets of near synonyms) and described by lexico-semantic relations linking synsets into a complex network.

In order to reduce the variety of ways for expressing identical lexical meanings text words can be mapped onto the appropriate synsets – synonymous word uses are mapped to the same synset. However, two problems appear. Such mapping requires the use of a Word Sense Disambiguation (henceforth WSD) tool that still expresses a significant error (around 30%). Moreover, not all words in text are used in their literal meaning and the same words can be used in utterances describing or referring to different subtopics. The same subtopic can be discussed in different texts with slightly different words. However, we assume that all words used for discussing the same subtopic are closely semantically related. Thus synsets corresponding to them are located in the same regions of the wordnet graph of lexico-semantic relations. In order to cope with both problems, we decided to represent meaning of every word in text by a set consisting of the corresponding synset and synsets linked to it by paths in the wordnet graph of the limited length. Finally, in practice, we limited the paths to single links in order to avoid introducing semantic noise. Due to the different character of semantic relations, the synsets they link express varied information about each other. This is modelled by weights assigned to synsets: 1 for the corresponding synset and < 1 for synsets linked by relations. So, finally, a text word occurrence is mapped on a collection of synsets assigned weights from $(0, 1]$. In addition to relations linking synsets, selected relations linking lexical units are also utilised, i.e. for a synset s corresponding to a text word w, all synsets including lexical units that are linked to one of the s members are also added to the collection of w.

We used the WSD tool (Kędzia et al., 2015) to assign plWordNet synsets to words. Its accuracy is about 52% for tests on a balanced set of word senses and around 68% for an average text sample. That is why we have considered and evaluated also models in which the $p = 30\%$ or $p = 100\%$ top scored synsets were used to build a collection for a word. If the p percent of top scored synsets are selected, each gets a weight equal to the normalised probability produced by the WSD tool. In this model, a text word is mapped on the sum of collections build for the top scored synsets according to the WSD tool. If there is no synset for a text word in plWordNet, e.g. in the case of proper names, then such word is mapped to a singleton collection including its lemma with the weight 1.

The final weight for a member of the collection C is calculated as following: $w_c = w_s * w_r * (1/|r \in C|)$ where w_s, a synset weight, depends on WSD or how often was s added to the collection, w_r depends on the relation due to which s was added to C and the last constituent reduces weights for more frequent relations.

Due to the limited accuracy of the WSD tool, we decided also to test a simpler model in which a word w is mapped to a collection built from lemmas directly linked to w by lexico-semantic relations including synonymy expressed by synsets. No WSD is applied, all synsets that w belongs to, marked as $S(w)$, are used to build one merged collection of lemmas, not synsets. The collection for w is built from all synset members of $s \in S(w)$, as well as synset members linked to any $s \in S(w)$ by a lexico-semantic relation. The way of calculating weights is identical to the one applied to synset collections.

Three collection types were defined:

CollHHM – only synonymy, hyper/hyponymy, and meronymy/holonymy are used for building collections,

CollV1 – all relations from **CollHHM** plus relations used for the automated wordnet expansion (Piasecki et al., 2013): type/instance, inter-register synonymy, femininity, markedness (diminutive, augmentative, young being), antonymy and converse

CollV2 is **CollV1** with antonymy and converse excluded.

Concerning the values for relation weights, we also followed the solution developed for automated wordnet expansion: synonymy 1.0, hypernymy 0.49, hyponymy 0.7, meronymy/holonymy 0.42, type 0.49, instance 0.7, femininity 0.7, inter-register synonymy 0.7, markedness (diminutive, augmentative, young being) 0.7, antonymy 0.28, converse 0.28. The values tend to be correlated with the amount of the information that the target of the relation link delivers about the source.

Weights of the collection members discussed so far are aimed at expressing semantic information concerning the word represented by the given collection. However, such weights tell a little about how good is the given element in discriminating different texts. Discriminability should be calculated on the basis of a representative collection of documents. In the case of synset collection this is not possible, as there are no WSD corpora for Polish. Thus, in the case of both types of collections: the element specificity is estimated by the IDF factor (Salton and McGill, 1986) calculated for the lemma of this element on a basis of a text corpus. The weights are multiplied by IDF before calculating similarity of vectors. For computing similarity of vectors we analysed several measures. The best results were achieved with: cosine, Jaccard and Dice measures.

3. Evaluation

There is no golden standard for similarity of Polish short texts. Instead, we applied two kinds of evaluation: comparison with human judgements about the similarity of sentences and evaluation by application in the QA system. In all tests, texts were preprocessed by segmentation (sentences, tokens) and morpho-syntactic tagging (lemmatisation, disambiguation).

3.1. Semantic Similarity of Sentences

First, we wanted to compare the similarity measure values generated for sentence pairs with human similarity judgements. The test set consisted of 50 sentences classified into 5 categories[1] on the basis of assessment done by volunteers, 609 answers with 12 scores per a sentence on average. The range of scores per sentence was quite large in many cases and the inter-annotator agreement had to be low. Thus, these data provide only some insights and we have cleaned them by calculating average, median and removing cases that were identified as statistical anomalies.

In order to calculate IDFs we used a corpus of: 91,446 documents, where about 70% of them come from Wikipedia, the rest was selected from Wikinews.

Comparison of the proposed similarity measures with human evaluation was performed with two evaluation metrics: Mean Square Error and the *Significant Error Rate*.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{X}_i)^2 \quad (1)$$

where X_i is the value of the semantic similarity produced by the tested measure, and $\hat{X}_i$ is the average or median from the scores assigned by evaluators.

Significant Error Rate is calculated as the number of cases in which the tested measure expressed the error value beyond the lack of agreement between human evaluators, i.e. the difference is greater then the standard deviation:

$$SER = size(\{X_i : i \in [1,n] \& |X_i - \hat{X}_i| < \sigma_i\}) \quad (2)$$

As a baseline for the similarity we used cosine measure for lemma-based TF.IDF representation of test sentences. The MSE of the baseline was 0.0926 in relation to the average and 0.0994 for median. SER was 25 and 24, respectively. So, while MSE is relatively small, the baseline was beyond the evaluator disagreement in half of the cases.

Selected best results from the evaluation are presented in Tab. 1. They were compared with the average of the human scores. Similar results were obtained in comparison with median, but MSE was slightly higher and the number of test cases overcoming the baseline with respect to SER was lower. In Tab. 1, we can notice that all types of collections achieved results better than the baseline. In all cases weighting of relations improved the performance. Collections based on larger number of relations, namely CollV1 and CollV2 are better than CollHHM producing shorter vectors. Collections based on WSD express lower error in the case of taking only top synset, but collections based on lemmas mostly performed better.

However, as the number of test sentence pairs was limited and the agreement between evaluators low, we should not go too far with conclusions. Generally, the proposed measures showed their potential beyond a typical TF.IDF-based cosine measure, that is commonly used. In order to check the influence of the proposed measures in a more reliable way, we will analyse in the next section its influence on the large scale Question Answering system for Polish.

[1] 0 – the sentences are about different topics, 1 – are not equivalent but on the same topic, 2 – not equivalent, but share some details, 3 – approximately equivalent but some important informations are different, 4 – equivalent, but differ in some minor details, 5 – equivalent, have the same meaning.

3.2. Selection for Question Answering

In Question Answering (QA) systems the answer to the user question expressed in natural language is found by comparing it with documents and next snippets from documents. The goal is to find the sentence or snippet that is most likely to include the answer. Next the answer is extracted. However we focused on the first part: using the text similarity measures in comparing questions with documents and sentences. We assumed that a good similarity measure would improve the selection process. The evaluation was based on the *Borsuk* QA system for Polish (Marcińczuk et al., 2013). To asses the results we applied Mean Reciprocal Rank (MRR), typical for QA evaluation:

$$MRR = \frac{1}{n}\sum_{i=1}^{N}\frac{1}{rank_i} \quad (3)$$

where n is the number of questions, N is the number of documents returned for a question (constant for all questions) and $rank_i$ is the rank of a document including the correct answer for the question i.

The closer is the right answer to the top of the ranking the higher MRR is. The maximum value 1 is achieved if for all questions the answers are returned as the first ones. Position changes in the top part of the ranking have significant influence on MRR value, while rank changes on further positions have very limited effect on MRR.

In all cases questions were compared with individual sentences with the help of the proposed measures. The whole QA process was evaluated in three *accuracy modes*: document, overlapping snippet and snippet exactly matching the answer. In the document mode, document score is maximum over the scores of its sentences. However, in the case of documents including the answer, it is not checked if the selected maximum score sentence really includes the answer. In the snippet modes, the ranking is based on sentence scores. Sentences are expanded to snippets of $\pm m$ sentences around the analysed sentence. The score of the snippet equals to the score of the central sentence. In the overlapping mode, it is enough for the analysed snippet to overlap with the snippet including the answer to be considered as a positive choice. In the exact mode, the analysed snippet must match the snippet marked as the answer.

200 questions have been randomly selected from *Czy wiesz* dataset (Marcińczuk et al., 2013) for the needs of evaluation. In all tests, questions were first processed by *Borsuk* and next for each question 50 top-ranked documents found in the searching step were returned.

The proposed measures were used to re-ranked documents and snippets. Questions and document sentences were represented as weighted collections of plWordNet synsets, see Sec. 2. As there is no large corpus mapped to synsets, IDF weights were calculated locally on the basis of the 50 returned documents only. So, the IDF values describe the local specificity of synsets that could introduce some accidental bias. The results are presented in Tab. 2.

Key	Coll.	Weight	WSD	Mean Square Error			Significant Error Rate		
				Cos	Dice	Jac	Cos	Dice	Jac
lemmas	CollV2	n	–	0,0916	0,0987	0,1464	27	25	21
lemmas	CollHHM	y	–	0,0813	0,0840	0,1446	29	29	15
synsets	CollV1	y	best	0,0790	0,0814	0,1400	30	28	16
synsets	CollV2	y	all	0,0868	0,0902	0,1406	27	27	16
synsets	CollHHM	y	best	0,0886	0,0918	0,1534	26	26	17
synsets	CollHHM	y	30%	0,0893	0,0926	0,1511	25	25	17

Table 1: Comparison of the similarity measures with the average of manual evaluation.

Key	Coll.	Weight	WSD	Documents	Overlapping			Exact		
					m=0	m=1	m=2	m=0	m=1	m=2
Cosine										
lemmas	CollV2	y	–	0.6092	0.2287	0.4490	0.4798	0.0698	0.3436	0.4611
lemmas	CollV1	**n**	–	0.5747	0.2173	0.4171	0.4512	0.0734	0.3244	0.4315
synsets	Coll-HHM	y	30%	0.6122	0.2217	0.4550	0.4831	0.0705	0.3536	0.4590
synsets	Coll-HHM	y	all	0.5883	0.2082	0.4410	0.4695	0.0637	0.3464	0.4495
synsets	CollV2	y	30%	0.6011	0.2195	0.4496	0.4784	0.0750	0.3513	0.4592
Jaccard										
lemmas	CollV2	y	–	0.5984	0.2118	0.4324	0.4632	0.0737	0.3297	0.4427
lemmas	CollV1	y	–	0.5971	0.2239	0.4375	0.4637	0.0698	0.3421	0.4512
synsets	CollV2	y	best	0.5594	0.2287	0.4311	0.4557	0.0744	0.3289	0.4367
Unweighted correction factor × Cosine										
synsets	Coll-HHM	y	30%	0,6231	0,4191	0,5059	0,5334	0,1889	0,3914	0,5308
synsets	CollV2	y	all	0,6676	0,4068	0,5253	0,5533	0,1821	0,4108	0,5457
Weighted correction factor × Cosine										
lemmas	CollV2	y	–	0,6345	0,3514	0,5177	0,5439	0,1305	0,4006	0,5353
synsets	CollV2	y	all	0,6583	0,3441	0,5471	0,5735	0,1246	0,4307	0,5635
Basic configuration of QA system *Borsuk*										
–	–	–	–	0,8380	0,5369	0,7376	0,7697	0,2334	0,5861	0,7647
Borsuk enhanced with the selected similarity measures										
synsets	CollV2	y(w=0.08)	30%	0,8473	0,5553	0,7688	0,7996	0,2407	0,6020	0,7946
synsets	CollV2	y(w=0.07)	all	0,8439	0,5521	0,7633	0,7958	0,2389	0,5988	0,7908
synsets	Coll-HHM	y(w=0.08)	30%	0,8427	0,5513	0,7642	0,7951	0,2406	0,5989	0,7901

Table 2: Evaluation of similarity measures for short texts in the application to Question Answering.

As *Borsuk* is based on the *Lucene* (McCandless et al., 2010), we tested also how the proposed measure can fit into the scheme of the *Lucene Practical Scoring Function*, that is a complex equation with several constituents. Following the unweighted correction factor: $coord = |q \cap s|/|q|$ we proposed proportion correction factor in order to decrease accidental similarity of questions to short sentences:

$$coord_w = \frac{2\sum_{i=1}^{n} q_i s_i}{\sum_{i=1}^{n} q_i^2} \quad (4)$$

where q_i and s_i are weights in the vector representations of, respectively, a question and sentence (from a document).

Similarity measure was multiplied by the correction factors in the second group of tests. Selected best results for different measure variants are presented in Tab. 2. The best MRR scores for whole documents without correction factor were above 0.6, i.e. the proper answer was mostly on the first or second position. Concerning the results for snippets, we can notice that the accuracy of selecting one sentence as the answer is low, but it is also the case of the whole QA system *Borsuk*. Moreover, answers are often expressed in whole paragraphs in documents. Cosine measure expressed better results than Jacquard and Dice (not shown in Tab.2). However, the advantage of the cosine measure was mostly due to better treatment of short sentences from the test documents.

Multiplication by the correction factor mostly improved the overall results. The increases were especially significant for the comparison of questions with sentences and text snippets. The best results were achieved for the broader versions of the similarity measures, i.e. expanding lemmas or synsets with large number of relations. The similarity measure represents the lexical component in the complex comparison while the correction factor expresses the search heuristic. It is worth to notice, that in all tests the collection CollV2 with only 'positive' relations, i.e. without antonymy and converse (a specific type of antonymy) produced better results than CollV1 including both 'noisy' relations. The variants based on the exact choice of WSD were worse, but in the case of 30% and 'all' synsets used for collection construction, the results of WSD are visible in the weights assigned to the collection elements.

On the basis of the tests, three measure configurations: ⟨synsets, Coll-HHM, 30%⟩, ⟨synsets, CollV2, 30%⟩, ⟨synsets, CollV2, all⟩ – were selected for the tests inside the full *Borsuk* system. In all cases $m = 2$ was set for extracting snippets and the weighted correction factor.

3.3. Inside QA System

The selected measures have been added to the QA system *Borsuk* as an additional knowledge source for ranking the potential answers. The goal was to check if the use of a similarity measure can improve the overall performance of *Borsuk*. The optimised values for *Borsuk* parameters (Marcińczuk et al., 2013) were applied. The same set of 200 questions were used. For each question only the 50 top scored documents were analysed.

Borsuk ranks documents and text snippets according to a complex measure defined as a linear combination of several individual measures. In order to included the proposed measure in the complex one several weight values were tested. The final values are provided together with results expressed by the enhanced *Borsuk* in Tab. 2.

The introduction of the proposed measure into *Borsuk* ranking improved MRR by 0.009 for documents and by 0.01-0.03 for text snippets. These differences, as well as differences for text snippets are statistically significant according to Wilcoxon test (Wilcoxon, 1945). The differences may seem small, but the baseline of the optimised *Borsuk* was high and the observed increase of MRR was caused by improved positions of documents and snippets in the top part of the ranking and minor drops in the further part of the ranking. Manual inspection of the results showed that in many cases the shift was from the more remote ranking positions to the top three.

4. Conclusions

Semantic similarity measures for short texts were proposed. They are based on the description of lexical meanings in terms of the lexico-semantic relations provided by plWordNet. Text words are mapped onto semantic representation that is similar for words of the similar meaning. Some of the proposed measures showed improvement in the Question Answering system that can be attributed to better performance in selecting document and text snippets. Comparison of the produced similarity values for sentence pairs showed better correlation than a baseline solution based on a commonly used vector model. Half of the proposed methods utilise results of the WSD tool and produced good results that were better than we could expect from the limited accuracy of the tool. With the use of a better WSD tool the performance of the similarity methods can be improved. A wide set of wordnet relations was applied, but still selection of the final set and optimal assignment of weights to relation links must be found.

Acknowledgments Work supported by the Polish Ministry of Education and Science, Project CLARIN-PL, the European Innovative Economy Programme project POIG.01.01.02-14-013/09, and by the EU's 7FP under grant agreement No. 316097 [ENGINE].

5. References

Achananuparp, Palakorn, Xiaohua Hu, and Xiajiong Shen, 2008. The evaluation of sentence similarity measures. In Il-Yeol Song, Johann Eder, and ThoManh Nguyen (eds.), *Data Warehousing and Knowledge Discovery*, volume 5182 of *LNCS*. Springer, pages 305–316.

Bär, Daniel, Chris Biemann, Iryna Gurevych, , and Torsten Zesch, 2012. UKP: Computing semantic textual similarity by combining multiple content similarity measures. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*. ACL.

Corley, Courtney and Rada Mihalcea, 2005. Measuring the semantic similarity of texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*. ACL.

Fellbaum, Christiane (ed.), 1998. *WordNet – An Electronic Lexical Database*. The MIT Press.

Kędzia, Paweł, Maciej Piasecki, and Marlena J. Orlińska, 2015. Word sense disambiguation based on large scale polish clarin heterogeneous lexical resources. *Cognitive Studies*, 14(To appear).

Kouylekov, Milen and Bernardo Magnini, 2005. Recognizing textual entailment with tree edit distance. In *Proceedings of the PASCAL RTE Challenge*.

Li, Yuhua, David McLean, Zuhair A. Bandar, James D O'shea, and Keeley Crockett., 2006. Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering*, 18(8):1138–1150.

Liu, Hongzhe and Pengfei Wang., 2014. Assessing text semantic similarity using ontology. *Journal of Software*, 9(2):490–497.

Marcińczuk, Michał, Adam Radziszewski, Maciej Piasecki, Dominik Piasecki, and Marcin Ptak, 2013. Open dataset for development of Polish Question Answering systems. In *Proceedings of 6th Language & Technology Conference LTC 2013*. Poznań.

Maziarz, Marek, Maciej Piasecki, Ewa Rudnicka, and Stan Szpakowicz, 2014. plWordNet as the Cornerstone of a Toolkit of Lexico-semantic Resources. In *Proceedings of the Seventh Global Wordnet Conference*.

McCandless, Michael, Erik Hatcher, and Otis Gospodnetic, 2010. *Lucene in Action*. Manning Publications.

Mihalcea, Rada, Courtney Corley, and Carlo Strapparava, 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *Proc. of the 21st Nat. Conf. on Artificial Intelligence, Vol.1*, AAAI'06. AAAI.

Piasecki, Maciej, Radosław Ramocki, and Michał Kaliński, 2013. Information spreading in expanding wordnet hypernymy structure. In *Proc. of the Int. Conf. Recent Advances in Natural Language Processing RANLP 2013*. Hissar, Bulgaria: INCOMA Ltd.

Pourgholamali, Fatemeh and Mohsen Kahani, 2012. Semantic role based sentence compression. In *2nd International eConference on Computer and Knowledge Engineering (ICCKE)*. IEEE Computer Society.

Punyakanok, V., D. Roth, and W. Yih, 2004. Mapping dependencies trees: An application to question answering. In *Proceedings of AI&Math*.

Salton, Gerard and Michael J. McGill, 1986. *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc.

Wilcoxon, Frank, 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.