Offspring from Reproduction Problems: What Replication Failure Teaches Us

Antske Fokkens and **Marieke van Erp**
The Network Institute
VU University Amsterdam
Amsterdam, The Netherlands
{a.s.fokkens,m.g.j.van.erp}@vu.nl

Marten Postma
Utrecht University
Utrecht, The Netherlands
martenp@gmail.com

Ted Pedersen
Dept. of Computer Science
University of Minnesota
Duluth, MN 55812 USA
tpederse@d.umn.edu

Piek Vossen
The Network Institute
VU University Amsterdam
Amsterdam, The Netherlands
piek.vossen@vu.nl

Nuno Freire
The European Library
The Hague, The Netherlands
nfreire@gmail.com

Abstract

Repeating experiments is an important instrument in the scientific toolbox to validate previous work and build upon existing work. We present two concrete use cases involving key techniques in the NLP domain for which we show that reproducing results is still difficult. We show that the deviation that can be found in reproduction efforts leads to questions about how our results should be interpreted. Moreover, investigating these deviations provides new insights and a deeper understanding of the examined techniques. We identify five aspects that can influence the outcomes of experiments that are typically not addressed in research papers. Our use cases show that these aspects may change the answer to research questions leading us to conclude that more care should be taken in interpreting our results and more research involving systematic testing of methods is required in our field.

1 Introduction

Research is a collaborative effort to increase knowledge. While it includes validating previous approaches, our experience is that most research output in our field focuses on presenting new approaches, and to a somewhat lesser extent building upon existing work.

In this paper, we argue that the value of research that attempts to replicate previous approaches goes beyond simply *validating* what is already known. It is also an essential aspect for *building upon* existing approaches. Especially when validation fails or variations in results are found, systematic testing helps to obtain a clearer picture of both the approach itself and of the meaning of state-of-the-art results leading to a *better insight* into the quality of new approaches in relation to previous work.

We support our claims by presenting two use cases that aim to reproduce results of previous work in two key NLP technologies: measuring WordNet similarity and Named Entity Recognition (NER). Besides highlighting the difficulty of repeating other researchers' work, new insights about the approaches emerged that were not presented in the original papers. This last point shows that reproducing results is not merely part of good practice in science, but also an essential part in gaining a better understanding of the methods we use. Likewise, the problems we face in reproducing previous results are not merely frustrating inconveniences, but also pointers to research questions that deserve deeper investigation.

We investigated five aspects that cause experimental variation that are not typically described in publications: **preprocessing** (e.g. tokenisation), **experimental setup** (e.g. splitting data for cross-validation), **versioning** (e.g. which version of WordNet), **system output** (e.g. the exact features used for individual tokens in NER), and **system variation** (e.g. treatment of ties).

As such, reproduction provides a platform for systematically testing individual aspects of an approach that contribute to a given result. What is the influence of the size of the dataset, for example? How does using a different dataset affect the results? What is a reasonable divergence between different runs of the same experiment? Finding answers to these questions enables us to better interpret our state-of-the-art results.

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1691–1701,
Sofia, Bulgaria, August 4-9 2013.

Moreover, the experiments in this paper show that even while strictly trying to replicate a previous experiment, results may vary up to a point where they lead to different answers to the main question addressed by the experiment. The WordNet similarity experiment use case compares the performance of different similarity measures. We will show that the answer as to which measure works best changes depending on factors such as the gold standard used, the strategy towards part-of-speech or the ranking coefficient, all aspects that are typically not addressed in the literature.

The main contributions of this paper are the following:

1) An in-depth analysis of two reproduction use cases in NLP

2) New insights into the state-of-the-art results for WordNet similarities and NER, found because of problems in reproducing prior research

3) A categorisation of aspects influencing reproduction of experiments and suggestions on testing their influence systematically

The code, data and experimental setup for the WordNet experiments are available at `http://github.com/antske/WordNetSimilarity`, and for the NER experiments at `http://github.com/Mvanerp/NER`. The experiments presented in this paper have been repeated by colleagues not involved in the development of the software using the code included in these repositories. The remainder of this paper is structured as follows. In Section 2, previous work is discussed. Sections 3 and 4 describe our real-world use cases. In Section 5, we present our observations, followed by a more general discussion in Section 6. In Section 7, we present our conclusions.

2 Background

This section provides a brief overview of recent work addressing reproduction and benchmark results in computer science related studies and discusses how our research fits in the overall picture.

Most researchers agree that validating results entails that a method should lead to the same overall conclusions rather than producing the exact same numbers (Drummond, 2009; Dalle, 2012; Buchert and Nussbaum, 2012, etc.). In other words, we should strive to *reproduce* the same answer to a research question by different means, perhaps by re-implementing an algorithm or evaluating it on a new (in domain) data set. *Replication* has a somewhat more limited aim, and simply involves running the exact same system under the same conditions in order to get the exact same results as output.

According to Drummond (2009) replication is not interesting, since it does not lead to new insights. On this point we disagree with Drummond (2009) as replication allows us to: 1) validate prior research, 2) improve on prior research without having to rebuild software from scratch, and 3) compare results of reimplementations and obtain the necessary insights to perform reproduction experiments. The outcome of our use cases confirms the statement that deeper insights into an approach can be obtained when all resources are available, an observation also made by Ince et al. (2012).

Even if exact replication is not a goal many strive for, Ince et al. (2012) argue that insightful reproduction can be an (almost) impossible undertaking without the source code being available. Moreover, it is not always clear where replication stops and reproduction begins. Dalle (2012) distinguishes levels of reproducing results related to how close they are to the original work and how each contributes to research. In general, an increasing awareness of the importance of reproduction research and open code and data can be observed based on publications in high-profile journals (e.g. Nature (Ince et al., 2012)) and initiatives such as myExperiment.[1]

Howison and Herbsleb (2013) point out that, even though this is important, often not enough (academic) credit is gained from making resources available. What is worse, the same holds for research that investigates existing methods rather than introducing new ones, as illustrated by the question that is found on many review forms 'how novel is the presented approach?'. On the other hand, initiatives for journals addressing exactly this issue (Neylon et al., 2012) and tracks focusing on results verification at conferences such as VLDB[2] show that this opinion is not universal.

A handful of use cases on reproducing or replicating results have been published. Louridas and Gousios (2012) present a use case revealing that source code alone is not enough for reproducing

[1] `http://www.myexperiment.org`

[2] `http://www.vldb.org/2013/`

results, a point that is also made by Mende (2010) who provides an overview of all information required to replicate results.

The experiments in this paper provide use cases that confirm the points brought out in the literature mentioned above. This includes both observations that a detailed level of information is required for truly insightful reproduction research as well as the claim that such research leads to better understanding of our techniques. Furthermore, the work in this paper relates to Bikel (2004)'s work. He provides all information needed in addition to Collins (1999) to replicate Collins' benchmark results. Our work is similar in that we also aim to fill in the blanks needed to replicate results. It must be noted, however, that the use cases in this paper have a significantly smaller scale than Bikel's.

Our research distinguishes itself from previous work, because it links the challenges of reproduction to what they mean for reported results beyond validation. Ruml (2010) mentions variations in outcome as a reason not to emphasise comparisons to benchmarks. Vanschoren et al. (2012) propose to use experimental databases to systematically test variations for machine learning, but neither links the two issues together. Raeder et al. (2010) come closest to our work in a critical study on the evaluation of machine learning. They show that choices in the methodology, such as data sets, evaluation metrics and type of cross-validation can influence the conclusions of an experiment, as we also find in our second use case. However, they focus on the problem of evaluation and recommendations on how to achieve consistent reproducible results. Our contribution is to investigate how much results vary. We cannot control how fellow researchers carry out their evaluation, but if we have an idea of the variations that typically occur within a system, we can better compare approaches for which not all details are known.

3 WordNet Similarity Measures

Patwardhan and Pedersen (2006) and Pedersen (2010) present studies where the output of a variety of WordNet similarity and relatedness measures are compared. They rank Miller and Charles (1991)'s set (henceforth "*mc-set*") of 30 word pairs according to their semantic relatedness with several WordNet similarity measures.

Each measure ranks the *mc-set* of word pairs and these outputs are compared to Miller and Charles (1991)'s gold standard based on human rankings using the Spearman's Correlation Coefficient (Spearman, 1904, ρ). Pedersen (2010) also ranks the original set of 65 word pairs ranked by humans in an experiment by Rubenstein and Goodenough (1965) (*rg-set*) which is a superset of Miller and Charles's set.

3.1 Replication Attempts

This research emerged from a project running a similar experiment for Dutch on Cornetto (Vossen et al., 2013). First, an attempt was made to reproduce the results reported in Patwardhan and Pedersen (2006) and Pedersen (2010) on the English WordNet using their WordNet::Similarity web-interface.[3] Results differed from those reported in the aforementioned works, even when using the same versions as the original, WordNet::Similarity-1.02 and WordNet 2.1 (Patwardhan and Pedersen, 2006) and WordNet::Similarity-2.05 and WordNet 3.0 (Pedersen, 2010), respectively.[4]

The fact that results of similarity measures on WordNet can differ even while the same software and same versions are used indicates that properties which are not addressed in the literature may influence the output of similarity measures. We therefore conducted a range of experiments that, in addition to searching for the right settings to replicate results of previous research, address the following questions:

1) Which properties have an impact on the performance of WordNet similarity measures?

2) How much does the performance of individual measures vary?

3) How do commonly used measures compare when the variation of their performance are taken into account?

3.2 Methodology and first observations

The questions above were addressed in two stages. In the first stage, Fokkens, who was not involved in the first replication attempt implemented a script to calculate similarity measures using WordNet::Similarity. This included similarity measures introduced by Wu and Palmer (1994) (`wup`),

[3]Obtained from `http://talisker.d.umn.edu/cgi-bin/similarity/similarity.cgi`, WordNet::Similarity version 2.05. This web interface has now moved to `http://maraca.d.umn.edu`

[4]WordNet::Similarity were obtained `http://search.cpan.org/dist/WordNet-Similarity/`.

Leacock and Chodorow (1998) (lch), Resnik (1995) (res), Jiang and Conrath (1997) (jcn), Lin (1998) (lin), Banerjee and Pedersen (2003) (lesk), Hirst and St-Onge (1998) (hso) and Patwardhan and Pedersen (2006) (vector and vpairs) respectively.

Consequently, settings and properties were changed systematically and shared with Pedersen who attempted to produce the new results with his own implementations. First, we made sure that the script implemented by Fokkens could produce the same WordNet similarity scores for each individual word pair as those used to calculate the ranking on the *mc-set* by Pedersen (2010). Finally, the gold standard and exact implementation of the Spearman ranking coefficient were compared.

Differences in results turned out to be related to variations in the **experimental setup**. First, we made different assumptions on the restriction of part-of-speech tags (henceforth "PoS-tag") considered in the comparison. Miller and Charles (1991) do not discuss how they deal with words with more than one PoS-tag in their study. Pedersen therefore included all senses with any PoS-tag in his study. The first replication attempt had restricted PoS-tags to nouns based on the idea that most items are nouns and subjects would be primed to primarily think of the noun senses. Both assumptions are reasonable. Pos-tags were not restricted in the second replication attempt, but because of a bug in the code only the first identified PoS-tag ("noun" in all cases) was considered. We therefore mistakenly assumed that PoS-tag restrictions did not matter until we compared individual scores between Pedersen and the replication attempts.

Second, there are two gold standards for the Miller and Charles (1991) set: one has the scores assigned during the original experiment run by Rubenstein and Goodenough (1965), the other has the scores assigned during Miller and Charles (1991)'s own experiment. The ranking correlation between the two sets is high, but they are not identical. Again, there is no reason why one gold standard would be a better choice than the other, but in order to replicate results, it must be known which of the two was used. Third, results changed because of differences in the treatment of ties while calculating Spearman ρ. The influence of the exact gold standard and calculation of Spearman ρ could only be found because Pedersen could provide the output of the similarity measures he used to calculate the coefficient. It is unlikely we would have been able to replicate his results at all without the output of this intermediate step. Finally, results for lch, lesk and wup changed according to measure specific configuration settings such as including a PoS-tag specific root node or turning on normalisation.

measure	Spearman ρ		Kendall τ		ranking
	min	max	min	max	variation
path based similarity					
path	0.70	0.78	0.55	0.62	1-8
wup	0.70	0.79	0.53	0.61	1-6
lch	0.70	0.78	0.55	0.62	1-7
path based information content					
res	0.65	0.75	0.26	0.57	4-11
lin	0.49	0.73	0.36	0.53	6-10
jcn	0.46	0.73	0.32	0.55	5, 7-11
path based relatedness					
hso	0.73	0.80	0.36	0.41	1-3,5-10
dictionary and corpus based relatedness					
vpairs	0.40	0.70	0.26	0.50	7-11
vector	0.48	0.92	0.33	0.76	1,2,4,6-11
lesk	0.66	0.83	-0.02	0.61	1-8,11,12

Table 1: Variation WordNet measures' results

In the second stage of this research, we ran experiments that systematically manipulate the influential factors described above. In this experiment, we included both the *mc-set* and the complete *rg-set*. The implementation of Spearman ρ used in Pedersen (2010) assigned the lowest number in ranking to ties rather than the mean, resulting in an unjustified drop in results for scores that lead to many ties. We therefore experimented with a different correlation measure, Kendall tau coefficient (Kendall, 1938, τ) rather than two versions of Spearman ρ.

3.3 Variation per measure

All measures varied in their performance. The complete outcome of our experiments (both the similarity measures assigned to each pair as well as the output of the ranking coefficients) are included in the data set provided at http://github.com/antske/WordNetSimilarity. Table 1 presents an overview of the main point we wish to make through this experiment: the minimal and maximal results according to both ranking coefficients. Results for similarity measures varied from 0.06-0.42 points for Spearman ρ and from 0.05-0.60 points for Kendall τ. The last column indicates the variation of performance of a measure

compared to the other measures, where 1 is the best performing measure and 12 is the worst.[5] For instance, `path` has been best performing measure, second best, eighth best and all positions in between, `vector` has ranked first, second and fourth, but also occupied all positions from six to eleven.

In principle, it is to be expected that numbers are not exactly the same while evaluating against a different data set (the *mc-set* versus the *rg-set*), taking a different set of synsets to evaluate on (changing PoS-tag restrictions) or changing configuration settings that influence the similarity score. However, a variation of up to 0.44 points in Spearman ρ and 0.60 in Kendall τ[6] leads to the question of how indicative these results really are. A more serious problem is the fact that the comparative performance of individual measure changes. Which measure performs best depends on the evaluation set, ranking coefficient, PoS-tag restrictions and configuration settings. This means that the answer to the question of which similarity measure is best to mimic human similarity scores depends on aspects that are often not even mentioned, let alone systematically compared.

3.4 Variation per category

For each influential category of experimental variation, we compared the variation in Spearman ρ and Kendall τ, while similarity measure and other influential categories were kept stable. The categories we varied include WordNet and WordNet::Similarity version, the gold standard used to evaluate, restrictions on PoS-tags, and measure specific configurations. Table 2 presents the maximum variation found across measures for each category. The last column indicates how often the ranking of a specific measure changed as the category changed, e.g. did the measure ranking third using specific configurations, PoS-tag restrictions and a specific gold standard using WordNet 2.1 still rank third when WordNet 3.0 was used instead? The number in parentheses next to the 'different ranks' in the table presents the total number of scores investigated. Note that this number changes for each category, because we compared two WordNet versions (WN version), three gold standard and PoS-tag restriction variations and configuration only for the subset of scores where configuration matters.

Variation	Maximum difference		Different
	Spearman ρ	Kendall τ	rank (tot)
WN version	0.44	0.42	223 (252)
gold standard	0.24	0.21	359 (504)
PoS-tag	0.09	0.08	208 (504)
configuration	0.08	0.60	37 (90)

Table 2: Variations per category

There are no definite statements to make as to which version (Patwardhan and Pedersen (2006) vs Pedersen (2010)), PoS-tag restriction or configuration gives the best results. Likewise, while most measures do better on the smaller data set, some achieve their highest results on the full set. This is partially due to the fact that ranking coefficients are sensitive to outliers. In several cases where PoS-tag restrictions led to different results, only one pair received a different score. For instance, `path` assigns a relatively high score to the pair *chord-smile* when verbs are included, because the hierarchy of verbs in WordNet is relatively flat. This effect is not observed in `wup` and `lch` which correct for the depth of the hierarchy. On the other hand, `res`, `lin` and `jcn` score better on the same set when verbs are considered, because they cannot detect any relatedness for the pair *crane-implement* when restricted to nouns.

On top of the variations presented above, we notice a discrepancy between the two coefficients. Kendall τ generally leads to lower coefficiency scores than Spearman ρ. Moreover, they each give different relative indications: where `lesk` achieves its highest Spearman ρ, it has an extremely low Kendall τ of 0.01. Spearman ρ uses the difference in rank as its basis to calculate a correlation, where Kendall τ uses the number of items with the correct rank. The low Kendall τ for `lesk` is the result of three pairs receiving a score that is too high. Other pairs that get a relatively accurate score are pushed one place down in rank. Because only items that receive the exact same rank help to increase τ, such a shift can result in a drastic drop in the coefficient. In our opinion, Spearman ρ is therefore preferable over Kendall τ. We included τ, because many authors do not mention the ranking coefficient they use (cf. Budanitsky and Hirst (2006), Resnik (1995)) and both ρ and τ are com-

[5] Some measures ranked differently as their individual configuration settings changed. In these cases, the measure was included in the overall ranking multiple times, which is why there are more ranking positions than measures.

[6] Section 3.4 explains why the variation in Kendall is this extreme and ρ is more appropriate for this task.

monly used coefficients.

Except for WordNet, which Budanitsky and Hirst (2006) hold accountable for minor variations in a footnote, the influential categories we investigated in this paper, to our knowledge, have not yet been addressed in the literature. Cramer (2008) points out that results from WordNet-Human similarity correlations lead to scattered results reporting variations similar to ours, but she compares studies using different measures, data and experimental setup. This study shows that even if the main properties are kept stable, results vary enough to change the identity of the measure that yields the best performance. Table 1 reveals a wide variation in ranking relative to alternative approaches. Results in Table 2 show that it is common for the ranking of a score to change due to variations that are not at the core of the method.

This study shows that it is far from clear how different WordNet similarity measures relate to each other. In fact, we do not know how we can obtain the best results. This is particularly challenging, because the 'best results' may depend on the intended use of the similarity scores (Meng et al., 2013). This is also the reason why we presented the maximum variation observed, rather than the average or typical variation (mostly below 0.10 points). The experiments presented in this paper resulted in a vast amount of data. An elaborate analysis of this data is needed to get a better understanding of how measures work and why results vary to such an extent. We leave this investigation to future work. If there is one take-home message from this experiment, it is that one should experiment with parameters such as restrictions on PoS-tags or configurations and determine which score to use depending on what it is used for, rather than picking something that did best in a study using different data for a different task and may have used a different version of WordNet.

4 Reproducing a NER method

Freire et al. (2012) describe an approach to classifying named entities in the cultural heritage domain. The approach is based on the assumption that domain knowledge, encoded in complex features, can aid a machine learning algorithm in NER tasks when only little training data is available. These features include information about person and organisation names, locations, as well as PoS-tags. Additionally, some general features are used such as a window of three preceding and two following tokens, token length and capitalisation information. Experiments are run in a 10-fold cross-validation setup using an open source machine learning toolkit (McCallum, 2002).

4.1 Reproducing NER Experiments

This experiment can be seen as a real-world case of *the sad tale of the Zigglebottom tagger* (Pedersen, 2008). The (fictional) Zigglebottom tagger is a tagger with spectacular results that looks like it will solve some major problems in your system. However, the code is not available and a new implementation does not yield the same results. The original authors cannot provide the necessary details to reproduce their results, because most of the work has been done by a PhD student who has finished and moved on to something else. In the end, the newly implemented Zigglebottom tagger is not used, because it does not lead to the promised better results and all effort went to waste.

Van Erp was interested in the NER approach presented in Freire et al. (2012). Unfortunately, the code could not be made available, so she decided to reimplement the approach. Despite feedback from Freire about particular details of the system, results remained 20 points below those reported in Freire et al. (2012) in overall F-score (Van Erp and Van der Meij, 2013).

The reimplementation process involved choices about seemingly small details such as rounding to how many decimals, how to tokenise or how much data cleanup to perform (normalisation of non-alphanumeric characters for example). Trying different parameter combinations for feature generation and the algorithm never yielded the exact same results as Freire et al. (2012). The results of the best run in our first reproduction attempt, together with the original results from Freire et al. (2012) are presented in Table 3. Van Erp and Van der Meij (2013) provide an overview of the implementation efforts.

4.2 Following up from reproduction

Since the experiments in Van Erp and Van der Meij (2013) introduce several new research questions regarding the influence of data cleaning and the limitations of the dataset, we performed some additional experiments.

First, we varied the tokenisation, removing non-alphanumeric characters from the data set. This yielded a significantly smaller data set (10,442

	(Freire et al., 2012) results			Van Erp and Van der Meij's replication results		
	Precision	Recall	$F_{\beta=1}$	Precision	Recall	$F_{\beta=1}$
LOC (388)	92%	55%	69	77.80%	39.18%	52.05
ORG (157)	90%	57%	70	65.75%	30.57%	41.74
PER (614)	91%	56%	69	73.33%	37.62%	49.73
Overall (1,159)	91%	55%	69	73.33%	37.19%	49.45

Table 3: Precision, recall and $F_{\beta=1}$ scores for the original experiments from Freire et al. 2012 and our replication of their approach as presented in Van Erp and Van der Meij (2013)

tokens vs 12,510), and a 15 point drop in overall F-score. Then, we investigated whether variation in the cross-validation splits made any difference as we noticed that some NEs were only present in particular fields in the data, which can have a significant impact on a small dataset. We inspected the difference between different cross-validation folds by computing the standard deviations of the scores and found deviations of up to 25 points in F-score between the 10 splits. In the general setup, database records were randomly distributed over the folds and cut off to balance the fold sizes. In a different approach to dividing the data by distributing individual sentences from the records over the folds, performance increases by 8.57 points in overall F-score to 58.02. This is not what was done in the original Freire et al. (2012) paper, but shows that the results obtained with this dataset are quite fragile.

As we worried about the complexity of the feature set relative to the size of the data set, we deviated somewhat from Freire et al. (2012)'s experiments in that we switched some features on and off. Removal of complex features pertaining to the window around the focus token improved our results by 3.84 points in overall F-score to 53.39. The complex features based on VIAF,[7] GeoNames[8] and WordNet do contribute to the classification in the Mallet setup as removing them and only using the focus token, window and generic features causes a slight drop in overall F-score from 49.45 to 47.25.

When training the Stanford NER system (Finkel et al., 2005) on just the tokens from the Freire data set and the parameters from english.all.3class.distsim.prop (included in the Stanford NER release, see also Van Erp and Van der Meij (2013)), our F-scores come very close to those reported by Freire et al. (2012), but mostly with a higher recall and lower precision. It is puzzling that the Stanford system obtains such high results with only very simple features, whereas for Mallet the complex features show improvement over simpler features. This leads to questions about the differences between the CRF implementations and the influence of their parameters, which we hope to investigate in future work.

4.3 Reproduction difficulties explained

Several reasons may be the cause of why we fail to reproduce results. As mentioned, not all resources and data were available for this experiment, thus causing us to navigate in the dark as we could not reverse-engineer intermediate steps, but only compare to the final precision, recall and F-scores.

The experiments follow a general machine learning setup consisting roughly of four steps: preprocess data, generate features, train model and test model. The novelty and replication problems lie in the first three steps. How the data was preprocessed is a major factor here. The data set consisted of XML files marked up with inline named entity tags. In order to generate machine learning features, this data has to be tokenised, possibly cleaned up and the named entity markup had to be converted to a token-based scheme. Each of these steps can be carried out in several ways, and choices made here can have great influence on the rest of the pipeline.

Similar choices have to be made for preprocessing external resources. From the descriptions in the original paper, it is unclear how records in VIAF and GeoNames were preprocessed, or even which versions of these resources were used. Preprocessing and calculating occurrence statistics over VIAF takes 30 hours for each run. It is thus not feasible to identify the main potential variations without the original data to verify this prepatory step.

Numbers had to be rounded when generating the features, leading to the question of how many decimals are required to be discriminative without creating an overly sparse dataset. Freire recalls that encoding features as multi-value discrete fea-

[7] http://www.viaf.org
[8] http://www.geonames.org

tures versus several boolean features can have significant impact. These settings are not mentioned in the paper, making reproduction very difficult.

As the project in which the original research was performed has ended, and there is no central repository where such information can be retrieved, we are left to wonder how to reuse this approach in order to further domain-specific NER.

5 Observations

In this section, we generalise the observations from our use cases to the main categories that can influence reproduction.

Despite our efforts to describe our systems as clearly as possible, details that can make a tremendous difference are often omitted in papers. It will be no surprise to researchers in the field that **preprocessing** of data can make or break an experiment.

The choice of which steps we perform, and how each of these steps is carried out exactly are part of our **experimental setup**. A major difference in the results for the NER experiments was caused by variations in the way in which we split the data for cross-validation.

As we fine-tune our techniques, software gets updated, data sets are extended or annotation bugs are fixed. In the WordNet experiment, we found that there were two different gold standard data sets. There are also different versions of WordNet, and the WordNet::Similarity packages. Similarly for the NER experiment, GeoNames, VIAF and Mallet are updated regularly. It is therefore critical to pay attention to **versioning**.

Our experiments often consist of several different steps whose outputs may be difficult to retrace. In order to check the output of a reproduction experiment at every step of the way, **system output** of experiments, including intermediate steps, is vital. The WordNet replication was only possible, because Pedersen could provide the similarity scores of each word pair. This enabled us to compare the intermediate output and identify the source of differences in output.

Lastly, there may be inherent **system variations** in the techniques used. Machine learning algorithms may for instance use coin flips in case of a tie. This was not observed in our experiments, but such variations may be determined by running an experiment several times and taking the average over the different runs (cf. Raeder et al. (2010)).

All together, these observations show that sharing data and software play a key role in gaining insight into how our methods work. Vanschoren et al. (2012) propose a setup that allows researchers to provide their full experimental setup, which should include exact steps followed in preprocessing the data, documentation of the experimental setup, exact versions of the software and resources used and experimental output. Having access to such a setup allows other researchers to validate research, but also tweak the approach to investigate system variation, systematically test the approach in order to learn its limitations and strengths and ultimately improve on it.

6 Discussion

Many of the aspects addressed in the previous section such as preprocessing are typically only mentioned in passing, or not at all. There is often not enough space to capture all details, and they are generally not the core of the research described. Still, our use cases have shown that they can have a tremendous impact on reproduction, and can even lead to different conclusions. This leads to serious questions on how we can interpret our results and how we can compare the performance of different methods. Is an improvement of a few per cent really due to the novelty of the approach if larger variations are found when the data is split differently? Is a method that does not quite achieve the highest reported state-of-the-art result truly less good? What does a state-of-the-art result mean if it is only tested on one data set?

If one really wants to know whether a result is better or worse than the state-of-the-art, the range of variation within the state-of-the-art must be known. Systematic experiments such as the ones we carried out for WordNet similarity and NER, can help determine this range. For results that fall within the range, it holds that they can only be judged by evaluations going beyond comparing performance numbers, i.e. an evaluation of how the approach achieves a given result and how that relates to alternative approaches.

Naturally, our use cases do not represent the entire gamut of research methodologies and problems in the NLP community. However, they do represent two core technologies and our observations align with previous literature on replication and reproduction.

Despite the systematic variation we employed

in our experiments, they do not answer all questions that the problems in reproduction evoked. For the WordNet experiments, deeper analysis is required to gain full understanding of how individual influential aspects interact with each measurement. For the NER experiments, we are yet to identify the cause of our failure to reproduce.

The considerable time investment required for such experiments forms a challenge. Due to pressure to publish or other time limitations, they cannot be carried out for each evaluation. Therefore, it is important to share our experiments, so that other researchers (or students) can take this up. This could be stimulated by instituting reproduction tracks in conferences, thus rewarding systematic investigation of research approaches. It can also be aided by adopting initiatives that enable authors to easily include data, code and/or workflows with their publications such as the PLOS/figshare collaboration.[9] We already do a similar thing for our research problems by organising challenges or shared tasks, why not extend this to systematic testing of our approaches?

7 Conclusion

We have presented two reproduction use cases for the NLP domain. We show that repeating other researchers' experiments can lead to new research questions and provide new insights into and better understanding of the investigated techniques.

Our WordNet experiments show that the performance of similarity measures can be influenced by the PoS-tags considered, measure specific variations, the rank coefficient and the gold standard used for comparison. We not only find that such variations lead to different numbers, but also different rankings of the individual measures, i.e. these aspects lead to a different answer to the question as to which measure performs best. We did not succeed in reproducing the NER results of Freire et al. (2012), showing the complexity of what seems a straightforward reproduction case based on a system description and training data only. Our analyses show that it is still an open question whether additional complex features improve domain specific NER and that this may partially depend on the CRF implementation.

Some observations go beyond our use cases. In particular, the fact that results vary significantly because of details that are not made explicit in our publications. Systematic testing can provide an indication of this variation. We have classified relevant aspects in five categories occurring across subdisciplines of NLP: **preprocessing**, **experimental setup**, **versioning**, **system output**, and **system variation**.

We believe that knowing the influence of different aspects in our experimental workflow can help increase our understanding of the robustness of the approach at hand and will help understand the meaning of the state-of-the-art better. Some techniques are reused so often (the papers introducing WordNet similarity measures have around 1,000-2,000 citations each as of February 2013, for example) that knowing their strengths and weaknesses is essential for optimising their use.

As mentioned many times before, sharing is key to facilitating reuse, even if the code is imperfect and contains hacks and possibly bugs. In the end, the same holds for software as for documentation: *it is like sex: if it is good, it is very good and if it is bad, it is better than nothing!*[10] But most of all: when reproduction fails, regardless of whether original code or a reimplementation was used, valuable insights can emerge from investigating the cause of this failure. So don't let your failing reimplementations of the *Zigglebottom* tagger collect dusk on a shelf while others reimplement their own failing *Zigglebottoms*. As a community, we need to know where our approaches fail, as much –if not more– as where they succeed.

Acknowledgments

We would like to thank the anonymous reviewers for their eye to detail and useful comments to make this a better paper. We furthermore thank Ruben Izquierdo, Lourens van der Meij, Christoph Zwirello, Rebecca Dridan and the Semantic Web Group at VU University for their help and useful feedback. The research leading to this paper was supported by the European Union's 7th Framework Programme via the NewsReader Project (ICT-316404), the Agora project, by NWO CATCH programme, grant 640.004.801, and the BiographyNed project, a joint project with Huygens/ING Institute of the Dutch Academy of Sciences funded by the Netherlands eScience Center (`http://esciencecenter.nl/`).

[9] `http://blogs.plos.org/plos/2013/01/easier-access-to-plos-data/`

[10] The documentation variant of this quote is attributed to Dick Brandon.

References

Stanjeev Banerjee and Ted Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 805–810, Acapulco, August.

Daniel M. Bikel. 2004. Intricacies of Collins' parsing model. *Computational Linguistics*, 30(4):479–511.

Tomasz Buchert and Lucas Nussbaum. 2012. Leveraging business workflows in distributed systems research for the orchestration of reproducible and scalable experiments. In Anne Etien, editor, *9ème édition de la conférence MAnifestation des JEunes Chercheurs en Sciences et Technologies de l'Information et de la Communication - MajecSTIC 2012 (2012)*.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

Michael Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Phd dissertation, University of Pennsylvania.

Irene Cramer. 2008. How well do semantic relatedness measures perform? a meta-study. In *Semantics in Text Processing. STEP 2008 Conference Proceedings*, volume 1, pages 59–70.

Olivier Dalle. 2012. On reproducibility and traceability of simulations. In *WSC-Winter Simulation Conference-2012*.

Chris Drummond. 2009. Replicability is not reproducibility: nor is it good science. In *Proceedings of the Twenty-Sixth International Conference on Machine Learning: Workshop on Evaluation Methods for Machine Learning IV*.

Jenny Finkel, Trond Grenager, and Christopher D. Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 363–370, Ann Arbor, USA.

Nuno Freire, José Borbinha, and Pável Calado. 2012. An approach for named entity recognition in poorly structured data. In *Proceedings of ESWC 2012*.

Graeme Hirst and David St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In C. Fellbaum, editor, *WordNet: An electronic lexical database*, pages 305–332. MIT Press.

James Howison and James D. Herbsleb. 2013. Sharing the spoils: incentives and collaboration in scientific software development. In *Proceedings of the 2013 conference on Computer Supported Cooperative Work*, pages 459–470.

Darrel C. Ince, Leslie Hatton, and John Graham-Cumming. 2012. The case for open computer programs. *Nature*, 482(7386):485–488.

Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics (ROCLING X)*, pages 19–33, Taiwan.

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

Claudia Leacock and Martin Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. In C. Fellbaum, editor, *WordNet: An electronic lexical database*, pages 265–283. MIT Press.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pages 296—304, Madison, USA.

Panos Louridas and Georgios Gousios. 2012. A note on rigour and replicability. *SIGSOFT Softw. Eng. Notes*, 37(5):1–4.

Andrew K. McCallum. 2002. MALLET: A machine learning for language toolkit. `http://mallet.cs.umass.edu`.

Thilo Mende. 2010. Replication of defect prediction studies: problems, pitfalls and recommendations. In *Proceedings of the 6th International Conference on Predictive Models in Software Engineering*. ACM.

Lingling Meng, Runqing Huang, and Junzhong Gu. 2013. A review of semantic similarity measures in wordnet. *International Journal of Hybrid Information Technology*, 6(1):1–12.

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

Cameron Neylon, Jan Aerts, C Titus Brown, Simon J Coles, Les Hatton, Daniel Lemire, K Jarrod Millman, Peter Murray-Rust, Fernando Perez, Neil Saunders, Nigam Shah, Arfon Smith, Gaël Varoquaux, and Egon Willighagen. 2012. Changing computational research. the challenges ahead. *Source Code for Biology and Medicine*, 7(2).

Siddharth Patwardhan and Ted Pedersen. 2006. Using wordnet based context vectors to estimate the semantic relatedness of concepts. In *Proceedings of the EACL 2006 Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together*, pages 1–8, Trento, Italy.

Ted Pedersen. 2008. Empiricism is not a matter of faith. *Computational Linguistics*, 34(3):465–470.

Ted Pedersen. 2010. Information content measures of semantic similarity perform better without sense-tagged text. In *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2010)*, pages 329–332, Los Angeles, USA.

Troy Raeder, T. Ryan Hoens, and Nitesh V. Chawla. 2010. Consequences of variability in classifier performance estimates. In *Proceedings of ICDM'2010*.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 448–453, Montreal, Canada.

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

Wheeler Ruml. 2010. The logic of benchmarking: A case against state-of-the-art performance. In *Proceedings of the Third Annual Symposium on Combinatorial Search (SOCS-10)*.

Charles Spearman. 1904. Proof and measurement of association between two things. *American Journal of Psychology*, 15:72—101.

Marieke Van Erp and Lourens Van der Meij. 2013. Reusable research? a case study in named entity recognition. CLTL 2013-01, Computational Lexicology & Terminology Lab, VU University Amsterdam.

Joaquin Vanschoren, Hendrik Blockeel, Bernhard Pfahringer, and Geoffrey Holmes. 2012. Experiment databases. *Machine Learning*, 87(2):127–158.

Piek Vossen, Isa Maks, Roxane Segers, Hennie van der Vliet, Marie-Francine Moens, Katja Hofmann, Erik Tjong Kim Sang, and Maarten de Rijke. 2013. Cornetto: a Combinatorial Lexical Semantic Database for Dutch. In Peter Spyns and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch Results by the STEVIN-programme*, number XVII in Theory and Applications of Natural Language Processing, chapter 10. Springer.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 133—138, Las Cruces, USA.