An analytical approach to similarity measure selection for self-training

Vincent Van Asch VINCENT.VANASCH@UA.AC.BE
CLiPS Research Centre, University of Antwerp, Prinsstraat 13, 2000 Antwerpen, Belgium

Walter Daelemans WALTER.DAELEMANS@UA.AC.BE
CLiPS Research Centre, University of Antwerp, Prinsstraat 13, 2000 Antwerpen, Belgium

Keywords: similarity measures, domain knowledge, machine learning, part-of-speech tagging, self-training

Abstract

We present a framework for investigating properties of similarity measures as a criterion for selecting the best-suited measure for a specific task, in this paper: corpus selection for self-training. We focus on the squared Pearson's correlation coefficient as the property to rank similarity measures. Self-training is an unsupervised domain adaptation technique, in which three corpora are involved. Especially, the choice of the unlabeled corpus can be important and we show that similarity measures can be helpful when selecting an unlabeled corpus. In addition, we found that the correlation coefficient between similarity and accuracy of a similarity measure can be used to select the most suitable similarity measure, but other properties of similarity measures do also play a role.

1. Introduction

We first give a definition of *similarity measure*, since it is a vague term. In the context of this paper, a similarity measure is any function that produces a real number when applied to two text corpora. The output of the function should never switch sign and when two corpora are more similar, the absolute value of the similarity measure should be smaller. Divergences, like the Kullback-Leibler divergence, can be used as similarity measure, but divergences are certainly not the only candidates. When the two corpora are from different domains, the similarity measure can be called a domain similarity measure.

Domain similarity measures have been used in different natural language processing (NLP) setups (Zhang & Wang, 2009; McClosky, 2010; Plank, 2011) and, in general, the best-suited similarity measure depends on the task and the specific function of the similarity measure. Also combinations of different similarity measures have been tried. Nevertheless, it remains unclear which properties of a good similarity measure are responsible for its superiority. Our hypothesis is that a limited set of relevant properties exists and, depending on the processing task, some of the properties become more important than others. If this point of view is correct, creating an overview of existing similarity measures and their ranking for the different properties, would liberate the researcher from having to try all similarity measures and all combinations of measures in order to find the most appropriate measure(s).

In this paper, we investigate one candidate property, namely the degree of linear correlation between the similarity between two corpora and the accuracy in a machine learning experiment, using one of the corpora as the training corpus and the other corpus as the test corpus. The incentive to focus on the linear correlation comes from a general observation in domain adaptation literature: the more the domains of the test and the training corpus resemble each other, the better the performance of a machine learner will be. In addition, it has been found that for part-of-speech tagging, the correlation between accuracy and similarity is indeed linear (Van Asch & Daelemans, 2010).

When the linear correlation is selected as the discriminative property for similarity measures, it is possible to define what *best-suited* signifies. In this isolated situation, the *best-suited* similarity measure is the measure

Appearing in *Proceedings of BENELEARN 2013*.

that exhibits the highest squared Pearson correlation coefficient, r^2 (Pearson, 1896). With circular reasoning, this means that performance is the best similarity measure, because in this case the linear correlation would be perfect. Indeed, the best way to find out the result of an experiment is by running that experiment, but running an experiment can be time-consuming or there may be no annotated data available to actually compute the performance. For this reason, similarity measures that can be more quickly computed and that do not require annotated data are investigated. The linear correlation of these measures will be less strong, but, in return, they come with annotation-independence.

The remainder of this paper consists of an overview of the related research (Section 2), definitions of the different similarity measures that are used (Section 3), a presentation of the machine learning task (Section 4), the concept of self-training and the performance indicator (Section 5), experimental results (Section 6). A final section contains the conclusions and perspectives.

2. Related research

Divergences are used in natural language processing in various situations ranging from feature selection and training corpus creation to measuring the similarity between two language models (Della Pietra et al., 1997; Lee, 2001; Gao et al., 2002; Daumé III & Marcu, 2006; Chen et al., 2009; Mansour et al., 2009; Zhang & Wang, 2009; McClosky, 2010; Moore & Lewis, 2010; Plank, 2011). Some of the divergences that are used are perplexity, Kullback-Leibler divergence, and the Rényi divergence.

It is possible to use the divergence as such, using its value to draw inferences about corpora (Verspoor et al., 2009; Biber & Gray, 2010), but the most interesting usages apply the divergence to a machine learning system. A good example of such an application is the prediction of parsing accuracy (Ravi et al., 2008).

Despite the fact that authors have shown that a divergence (Van Asch & Daelemans, 2010; Plank, 2011) or a linear combination of divergences (McClosky, 2010) can be successfully used to link the similarity between domains to the performance of a natural language processing system, no consensus exists about which divergence or combination of divergences is best suited for the task. The best divergence is not selected on theoretical grounds but by testing a range of divergences and selecting the best one. Although this is a valid working method, in this paper we investigate if it is possible to select the best measure for a given task using the correlation of the divergence with the performance.

3. Similarity measures

A text corpus needs to be converted into a measurable representation if the goal is to express similarity between two corpora by means of a single figure. Examples of such representations are: a single figure (e.g. the average sentence length in the corpus) or a distribution (e.g. the relative frequencies of the unique tokens in a corpus). Similarity can be expressed by the difference between two representations or between combinations of representations. In this paper, we use similarity measures that are based on distributions, which are simple, yet expressive, representations of a corpus. A distribution P can be described formally as:

$$P = \left\{ p_k : p_k \in \mathbb{R}^+ \wedge \sum_i^n p_i = 1 \right\} \tag{1}$$

with $k \in \mathbb{N}$, a unique identifier for each unique token (=type), with p_k the relative frequency of a type k in the corpus, and n the number of unique tokens in the text corpus. Based on these distributions, the following similarity measures are tested in this paper: Kullback-Leibler divergence, KL (Kullback & Leibler, 1951), Rényi divergence, R (Rényi, 1961), Skew divergence, S (Lee, 1999), Jensen-Shannon divergence, JS (Lin, 1991), simple Unknown Word Ratio, sUWR (Zhang & Wang, 2009), and overlap. Overlap is the conceptual complement to sUWR.

Given two distributions: P based on a test corpus T and Q based on a training corpus S, the formulas of the similarity measures are:

$$KL(P;Q) = \sum_k p_k log_2\Big(\frac{p_k}{q_k}\Big) \tag{2}$$

$$R(P;Q;\alpha) = \frac{1}{(\alpha-1)} log_2\Big(\sum_k p_k^\alpha q_k^{1-\alpha}\Big) \text{ with } \alpha \geq 0 \tag{3}$$

$$S(P;Q) = KL\big(Q;\alpha P + (1-\alpha)Q\big) \text{ with } \alpha \in [0,1] \tag{4}$$

$$JS(P;Q) = \frac{1}{2}\Big[KL\Big(P;\frac{P+Q}{2}\Big) + KL\Big(Q;\frac{P+Q}{2}\Big)\Big] \tag{5}$$

$$sUWR = \frac{|\{k : p_k \neq 0 \wedge q_k = 0\}|}{|\{k : p_k \neq 0\}|} \tag{6}$$

$$Overlap = \frac{|\{k : p_k = 0 \wedge q_k \neq 0\}|}{|\{k : q_k \neq 0\}|} \tag{7}$$

With p_k the relative frequency of type k in corpus P, q_k the relative frequency of type k in corpus Q. If a type is not present in a distribution, it adopts a relative probability of 0.[1]

The measures have been chosen based on their suitability in tasks such as parsing and part-of-speech tagging (Lee, 2001; Daumé III & Marcu, 2006; Zhang & Wang, 2009; Van Asch & Daelemans, 2010; Plank, 2011) and overlap is chosen because it is an unsuitable measure. Overlap measures the proportion of types present in the training corpus, but not included in the test corpus. It is clear that this information is not necessarily helpful for predicting accuracy. The purpose is to have a similarity measure that deviates from the others.

4. NLP machine learning task

4.1. British National Corpus

The corpus that is used for the experiments is the British National corpus, BNC (BNC, 2001). This corpus contains part-of-speech labels and is divided into different domains.

Table 1. Overview of number of tokens and sentences in each domain of the BNC.

DOMAIN	# TOKENS	# SENTENCES
IMAGINATIVE	19,507,596	1,333,450
WORLD AFFAIRS	17,925,728	726,881
SOCIAL SCIENCE	13,481,239	542,410
LEISURE	11,088,447	560,094
ARTS	7,182,257	303,019
APPLIED SCIENCE	7,154,185	312,948
COMMERCE & FINANCE	6,787,847	302,455
NATURAL & PURE SCIENCE	4,095,326	172,836
BELIEF & THOUGHT	3,160,642	136,366

The BNC annotators provided nine domain codes (*i.e.* wridom codes), making it possible to divide the text from books and periodicals into nine subcorpora. These annotated semantic domains are: imaginative (wridom1), natural & pure science (wridom2), applied science (wridom3), social science (wridom4), world affairs (wridom5), commerce & finance (wridom6), arts (wridom7), belief & thought (wridom8), and leisure (wridom9). The smallest domain is the belief & thought domain, consisting of ∼3M tokens, see Table 1. To eliminate the influence of different corpus sizes, a random selection of approximately 1,500,000 tokens has been taken from each domain. During sampling, sentences are kept intact.

[1]For the Kullback-Leibler divergence, if $p_k \neq 0$ but $q_k = 0$, smoothing is applied, such that $q_k = 2^{-52}$.

4.2. Part-of-speech tagging

In this paper, we have chosen the part-of-speech tagging machine learning task, because of the substantial influence of domain differences on the performance for this task. The machine learner that is used for the experiments is the memory-based part-of-speech-tagger, MBT (Daelemans & van den Bosch, 2005). MBT[2] is a machine learner that stores examples in memory and uses an extension of the kNN algorithm to assign part-of-speech labels. The default settings were used. An advantage of MBT is its speed, making it the machine learner of choice to carry out a high number of experiments. In addition, the conclusions of this paper do not hinge upon the choice of the machine learner, since the linear correlation between similarity measure and accuracy is observed for other machine learners (Van Asch, 2012).

5. Self-training setup

5.1. Procedure

Self-training is a technique consisting of automatically labeling additional training data in a semi-supervised way, before running an experiment (Charniak, 1997; McClosky, 2010; Sagae, 2010). Jiang and Zhai (2007) present an example for part-of-speech tagging.

Three corpora are needed for self-training: a labeled, training corpus, a labeled test corpus, and an unlabeled additional corpus. During self-training, a model is learned from the training data and it is applied to the unlabeled data. Thus, the additional training data is created by automatically labeling unlabeled data. Next, the (partially incorrectly) labeled, additional data is appended to the original training data (*self-training step 1*). This first labeling step is followed by a second training phase. The model resulting from this phase is then used to label the test data (*self-training step 2*).

It remains under debate whether self-training is a useful method; it is not shown to lead to performance gain in every experimental setup. Sagae (2010) argues that self-training is only beneficial in those situations where the training and test data are sufficiently dissimilar, but other factors – such as labeling accuracy of the unlabeled data – have an influence too. It would be helpful if the positive effect of the application of self-

[2]Available at http://ilk.uvt.nl/mbt (Last accessed: March 2013)

training could be determined in advance. Thus, given a set of three corpora, the experimental question is: *Does a given setup lead to an accuracy increase when self-training is applied?*

5.2. Evaluation and performance indicator

F-score[3] can be used for evaluating the setups (van Rijsbergen, 1975). A *true positive (tp)* is a three-corpus setup that results in an accuracy increase and that has been predicted to benefit from self-training. A *false positive (fp)* is a setup that does not benefit from self-training, although it was predicted to do so. A *false negative (fn)* is a setup that benefits from self-training, but was predicted the converse.

For the experiments of this paper, when each setup is predicted to lead to accuracy gain, the F-score would be only 25.61% (see Section 6.2). This baseline is an indication of the general success of self-training. If self-training would always be helpful, this baseline would be 100%. But since this is not the case, the low baseline is an incentive to look for a way to predict whether self-training will be increasing performance or not for a given combination of corpora. To this end, a performance indicator δ is designed.

In our design, the performance indicator is a binary indicator: If the performance indicator is positive for a given setup, self-training is considered to be beneficial. If the indicator is negative, no gain is to be expected.

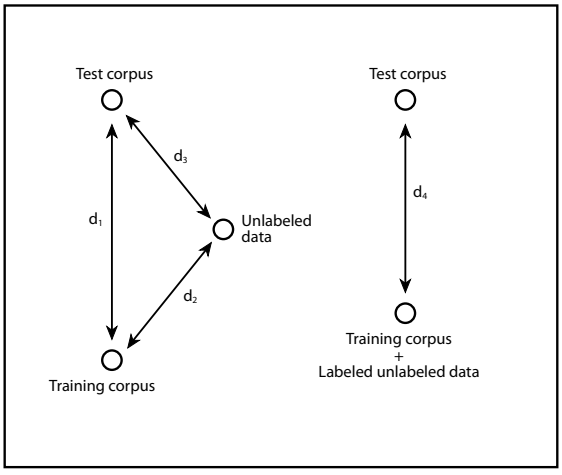


Figure 1. Theoretical justification of the performance indicator δ: Overview of similarities.

Figures 1 and 2 illustrate the rationale behind the design of the performance indicator δ. Figure 1 shows the different similarities that can be measured. d_1 represents the similarity between the training corpus and the test corpus. This is the only similarity involved when a straightforward test/train experiment is run. d_2 is the similarity between the training corpus and the additional unlabeled data. The labeling accuracy in the first self-training step is correlated with this similarity. d_3 is the similarity between the additional data and the test corpus. The more similar the test corpus and the additional data, the more beneficial self-training will be, provided that labeling during the first self-training step is near perfect. d_4 is the similarity between the composite corpus of training and additional data on the one side and the test data on the other. When labeling during the first self-training step would be perfect, the proportionality between d_4 and its associated accuracy (acc_4) would be the same as between d_1 and its accuracy (acc_1), since there would be no conceptual difference: both are measured between perfectly labeled corpora.

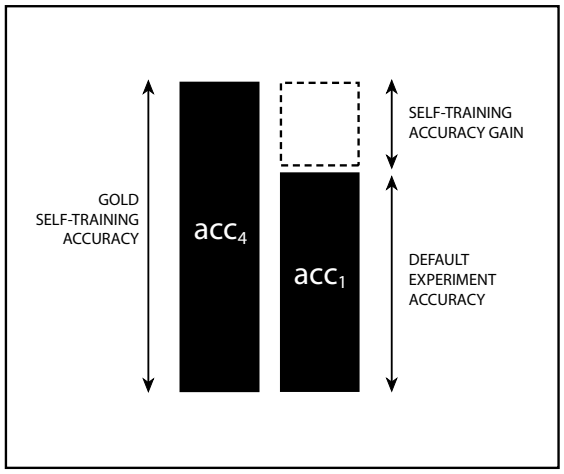


Figure 2. Theoretical justification of the performance indicator δ: Self-training accuracy gain.

It is known that $accuracy \propto \frac{1}{similarity}$ (Van Asch & Daelemans, 2010).[4] In a first step, the most important similarities are the similarity between *training corpus* and *test corpus* (d_1) and the similarity between *training corpus + additional data* and *test corpus* (d_4). The right column in Figure 2 depicts the accuracy of a regular test/train experiment (acc_1), and the height of this column is inverse proportional to the similarity d_1. Consider the case when labeling is perfect during the labeling step of a self-training experiment. In this case, the left column of Figure 2 is the highest obtain-

[3] F-score $= \frac{(1+\beta^2)\ tp}{(1+\beta^2)\ tp+\beta^2\ fp+fn}$; In this paper, β is set to 1.

[4] In this interpretation, the similarity value should be smaller when corpora are more alike.

able accuracy with self-training (acc_4). The perfectly labeled composite corpus serves as the training corpus. More data often leads to a higher performance e.g. Daelemans et al. (1999) and for that reason the left column is made higher than the right column.

The difference between acc_4 and acc_1 is the dashed column, which is the gain, obtained with (perfect) self-training, over a regular experiment. The performance indicator can be defined as

$$\delta' = \frac{acc_4}{acc_1} \tag{8}$$

if δ is larger than 1, self-training gain can be expected; if δ is smaller than 1, no gain is expected from self-training. Since we want to predict performance gain without running experiments, the accuracies are not available, but it is possible to use the similarities instead. In addition, the similarity between the *unlabeled data* and the *test data* (d_3) can be used as a proxy for d_4. Rewriting the performance indicator such that its outcome is binary then yields:

$$\delta = \frac{\left|\frac{d_1}{d_3} - 1\right|}{\frac{d_1}{d_3} - 1} \tag{9}$$

If δ is +1, gain is expected; if δ is -1, no gain is expected. The predictive power of this performance indicator is tested for part-of-speech self-training experiments in the next section.

6. Experiments

The corpus, the experimental setup, the evaluation method and the performance indicator have been presented in the previous section. In this section, these elements are used to conduct the experiments. First, the correlation coefficient for the different similarity measures is retrieved. Next, the self-training experiments are discussed.

6.1. Correlation r^2

The British national corpus contains nine domains, making it possible to select $\binom{9}{2} = 36$ different combinations of domains. The sets are used to conduct straightforward test/train experiments. Since it makes a difference whether a domain is selected as the first, i.e. as training corpus, or as the second, i.e. as test corpus, $36 \cdot 2! = 72$ experiments can be run.

By running the 72 part-of-speech tagging experiments, it is possible to compute the r^2 between the similarity measures between the test and training corpus on the one hand and the accuracy of the experiment on the other. In practice, each of the 72 experiments is a 25-fold cross-validation experiment. The training corpus is divided into five equal parts and the same is done for the test corpus. Next each training part is combined once with each test part in a part-of-speech tagging experiment with MBT. The final part-of-speech tagging accuracy and the similarity value are the averages of this cross-validation setup.

These experiments can be run while varying the similarity measure. The different correlations that are obtained in this manner will be used to differentiate the better from the worse similarity measures.

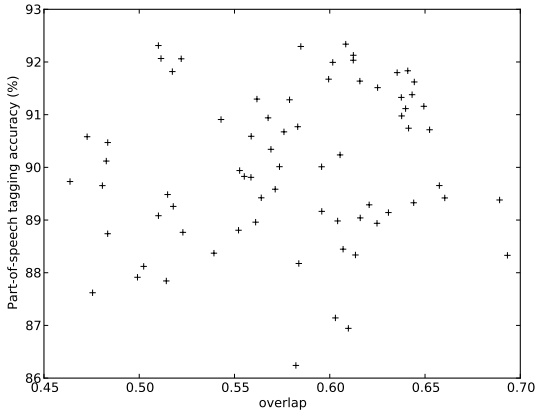
Table 2. The r^2 correlation coefficients for different similarity measures. The correlation is computed between similarity value and accuracy for 72 part-of-speech tagging experiments.

Measure	r^2
Rényi	0.083 – 0.987
Kullback-Leibler	0.986
Skew	0.224 – 0.985
sUWR	0.874
Jensen-Shannon	0.863
Overlap	0.051

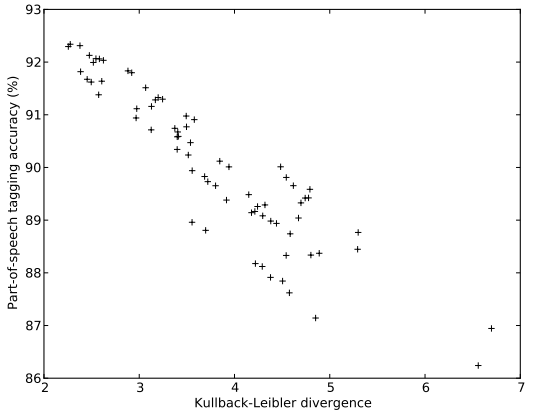
Table 2 shows the correlation coefficients r^2 for the selected set of similarity measures. Since the Skew and Rényi divergence contain a parameter α, a range of correlation coefficients is reported for these similarity measures.

The asymmetry of a measure M is the property that makes that the order of the distributions has an influence: $M(P,Q) \neq M(Q,P)$. Because all measures but the Jensen-Shannon divergence are asymmetric, the reported r^2 is an average value. For each run of 72 experiments, nine correlations are computed. One correlation for each set of 8 experiments for which the test corpus is the same. Averaging these values gives the values of Table 2. If all 72 experiments would be used to calculate a single overall r^2, the value for the Jensen-Shannon divergence would be too low since this measure cannot accommodate to the asymmetry of a part-of-speech tagging experiment: the similarity value is the same for $JS(P,Q)$ and $JS(Q,P)$, but the accuracy will be different. Splitting the computation of r^2 into a separate r^2 associated with each different test corpus, overcomes this incongruence, since $JS(P,Q)$ and $JS(Q,P)$ are no longer used for the calculation of the same r^2. Averaging all r^2's will aggregate the separate correlations to a single number.

As can be seen in Table 2, all measures show a good correlation except for overlap, which has been included for contrast. Two examples plots are given in Figure 3, along with the associated average r^2 value.



(a) Overlap ($r^2 = 0.051$)



(b) Kullback-Leibler ($r^2 = 0.986$)

Figure 3. Plot of two correlations between similarity value and part-of-speech accuracy for 72 experiments.

The parameterized divergences can also be adapted in such manner that they perform better or worse. The Rényi divergence has been applied with α varying from 0.02 to 0.98 in steps of 0.02. The higher α, the better the correlation. The Skew divergence with α values varying from 0.02 to 0.98 in steps of 0.02, varying from 0.9805 to 0.9995 in steps of 0.0005, and varying from 0.9995005 to 0.9999995 in steps of 5.10^{-7}. The higher α, the lower the correlation. Because the correlation of the Skew divergence declines much slower than the correlation of the Rényi divergence, more and smaller steps are computed for the Skew divergence.

6.2. Self-training gain prediction

The British National Corpus consists of nine domains and a set of three different corpora is needed to carry out a self-training experiment. This means that there are $\binom{9}{3} = 84$ possible sets. Since the order is important, and there are 3! permutations possible per set. In the end, this adds up to 504 experimental setups, using each domain either as training data, test data, or additional data.

As mentioned in Section 5.2, the baseline F-score when each self-training setup is expected to be beneficial is 25.61%. It should be stressed that a whole set of self-training setups are tested in this paper. As the baseline indicates, self-training may help performance, but it is not guaranteed. When examining a self-training setup for a single run of natural language processing task, one should be aware of the fact that a positive (or negative) outcome may be attributed to the corpora that have been selected. The single outcome should not give rise to conclusions about the general usability of the self-training technique for that task.

In this paper, when the 504 setups are tested during self-training experiments, only 74 experiments benefit from self-training. Leading to an F-score of $\frac{2 \cdot 74}{2 \cdot 74 + 430 + 0} = 25.61\%$.

When self-training is beneficial, the average performance gain is 0.07%, which amounts to an absolute difference of ∼985 tokens that are labeled correctly thanks to self-training. When self-training is harmful, the average performance loss is 0.09% or an extra of ∼1284 incorrectly labeled tokens. Overall, self-training has only a minor influence on accuracy, but even this minor influence can be predicted as is shown in the following experiments.

The derivation of the performance does not put severe constraints on the similarity measure that needs to be incorporated. The only requirements being that the value of the measure should never switch sign and that more similarity should lead to a smaller value. The 504 experiments are repeated while the similarity measure is replaced by one of similarity measures that are presented in Section 3.

We run a full round of experiments for the following similarity measures: Jensen-Shannon, Kullback-Leibler, sUWR, overlap, Rényi divergence with α varying from 0.02 to 0.98 in steps of 0.02, and Skew divergence with α varying from 0.02 to 0.98 in steps of 0.02, varying from 0.9805 to 0.9995 in steps of 0.0005, and varying from 0.9995005 to 0.9999995 in steps of

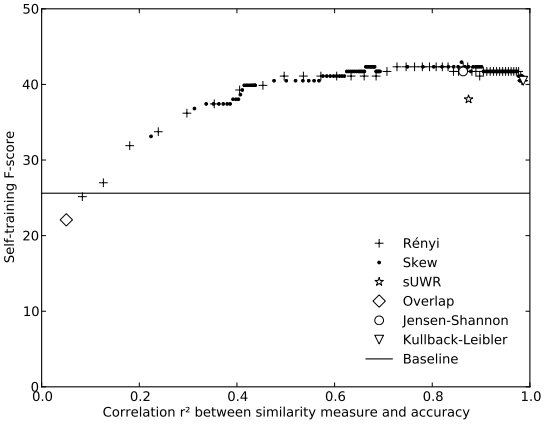


Figure 4. The variance of the self-training F-score for similarity measures that show different r^2 correlations for the accuracy of straightforward part-of-speech labeling experiments and the degree of test/train-similarity as expressed by that measure. Each point is a different measure. For the parameterized measures (Rényi and Skew), each point is the measure with a different α value, $\alpha \in]0, 1[$. For the Rényi divergence, an increasing α leads to a higher r^2. For the Skew divergence, an increasing α leads to a lower r^2.

0.0000005.[5] The F-scores of these experiments are plotted in Figure 4 and the F-scores are given in Table 3. The y-axis indicates the F-score, the x-axis indicates the correlation of the used measure, as reported in Table 2.

Table 3. F-scores for different similarity measures when used in the performance indicator δ. Statistical difference with baseline is indicated with *.

Measure	F-score
Rényi	25.15 – 42.33*
Kullback-Leibler	40.49*
Skew	33.13* – 42.94*
sUWR	38.04*
Jensen-Shannon	41.72*
Overlap	22.09

A first conclusion that can be drawn from Table 3 and Figure 4 is that it is almost always better to use the performance indicator to predict whether a self-training setup will be beneficial than to assume that self-training is beneficial. Only the two similarity measures at the left of the figure fall below the previously reported baseline of 25.61%. These two are overlap ($r^2 = 0.051$) and Rényi with $\alpha = 0.02$ ($r^2 = 0.083$).

The second conclusion is that, in general, similarity measures that are better correlated with accuracy (higher r^2) are more suited to be used as the core of the performance indicator δ. Although this observation holds in general, Figure 4 also shows that there is a broad range of similarity measures that approach the maximum F-score, provided that a certain degree of correlation has been reached. It is even the case that prediction appears to be less trustworthy when the higher r^2 scores are reached. The best Skew divergence is with $\alpha = 0.82$, associated with an r^2 value of 0.861 and reaching an F-score of 42.94%. Although a feeble downward trend for the top r^2 values can be observed, there is no statistical difference[6] between e.g. sUWR and Jensen-Shannon ($p = 0.099$). Only a larger difference, like between Jensen-Shannon and overlap ($p = 9.10^{-6}$), is statistically significant.

A higher r^2 does not necessitate obtaining a higher F-score. This fact can also be illustrated when straightforward accuracy is used as the similarity measure. As mentioned before, the best way to predict the accuracy of an experiment is by running that experiment. We can derive a similarity measure from the accuracy of an experiment: $similarity\ value = \frac{1}{accuracy}$. It is clear that the correlation r^2, computed as in Table 2, for this measure is 1. This *perfect* similarity measure can now be used in the performance indicator δ. The associated F-score for self-training then becomes 40.49%, which is not markedly better than using any other efficient measure. Since there is no better similarity measure available, this figure can be considered as a limitation to the method of using correlation r^2 as the selection criterion for selecting the best measure for this task. Indeed, if r^2 would be the only factor into play, the F-score when accuracy is used as similarity measure should be highest. But this is not the case.

This observation has consequences on two levels. First, r^2 cannot be used as the single selection criterion for selecting the best measure to be used in the performance indicator, although a minimal r^2 value is required. Second, the design of the performance indicator may not be flexible enough to anticipate certain situations, such as a very unsuccessful first labeling step. This implies that, even if you have built in the best similarity measure, it remains impossible to correctly predict all experimental setups for which self-training

[5]Because of the large amount of data points for the Skew divergence, with relatively small correlation differences, not all points are shown in Figure 4.

[6]Stratified approximate randomization testing of F-score of the positive class has been used to assess the significance of different labeling scores of the test set (Noreen, 1989). Implementation: www.clips.ua.ac.be/scripts/art

is beneficial.

6.3. Influence of the α parameter

When examining the definitions of the Rényi and Skew divergence, eqs. (3) and (4), we can draw the following conclusions on the influence of the α parameter on the measure: For the Rényi divergence, it can be seen that lowering α implies lowering the influence of the test corpus (p_k^α becomes smaller, and $q_k^{1-\alpha}$ becomes larger). As can be seen in Figure 4 by moving from right to left, lowering the influence of the test corpus leads to a deteriorated performance of the similarity measure, after a small initial gain.

For the Skew divergence, lowering alpha also means lowering the influence of the test corpus. Moving from left to right in Figure 4, in the beginning, lowering the influence of the test corpus improves the performance of the similarity measure, but when an α value of 0.82 is reached, the best parameter setting is reached. Further lowering of the influence of the test corpus will eventually lead to performance decrease.

Conclusion and perspectives

In this paper we investigated the possibility to rank similarity measures according to appropriateness for self-training. Our approach offers an analytical and systematic method to select the best-suited similarity measure from a set of measures. This, in contrast to the more frequent practical approach of testing all similarity measures in order to find the measure fit for the task. An additional advantage of the framework is that it enables the investigation of other properties besides linear correlation. This may be a stimulus for further research focusing on objective ways to express domain differences between corpora.

The machine learning task we implemented, is a self-training part-of-speech tagging task, in which a similarity measure is used to obtain a prediction about the usefulness of the self-training setup. To predict the usefulness, a performance indicator δ has been designed. We found that the r^2 of a similarity measure can be used as a coarse selection criterion for selecting a set of suitable measures.

The fact that the correlation cannot be used to single out one best-suited measure can be attributed to two interfering causes. The first cause being that the correlation coefficient may disregard certain influential properties of similarity measures. The sensitivity to relative frequency differences or the interdependencies between tokens may be two of such undetected properties. A second cause making the correlation appear an insufficient selection criterion may be that the effectiveness of the performance indicator δ can be limited by its design. This last conclusion is corroborated by the observation that incorporating a *perfect* similarity measure (accuracy) does not lead to the best performance.

For parameterized similarity measures (Rényi and Skew divergence), we found that moderately lowering the influence of the test corpus in the measure leads to an increased performance. This observation may contribute to the design of parameterized variants of existing similarity measures (like e.g. a parameterized sUWR). The newly introduced parameter should regulate the proportional influence of test and training corpus.

In general, we can conclude that the use of similarity measures in natural language processing is mainly a trial-end-error approach. We made start at looking into the various properties of similarity measures by investigating the information carried by the correlation coefficient. But, as our research showed, other properties exist and following research could focus on conceiving new measures that can express these properties in an objective manner.

Acknowledgements

This research is funded by the Research Foundation Flanders (FWO-project G.0478.10 – Statistical Relational Learning of Natural Language) and made possible through financial support from the University of Antwerp (GOA project BIOGRAPH).

References

Biber, D., & Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes*, *9*, 2–20.

BNC (2001). The British National Corpus, version 2 (BNC world). Available at http://www.natcorp.ox.ac.uk (Last accessed: March 2013).

Charniak, E. (1997). Statistical parsing with a context-free grammar and word statistics. *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Innovative Applications of Artificial Intelligence Conference* (pp. 598–603). Rhode Island, USA: MIT Press.

Chen, B., Lam, W., Tsang, I., & Wong, T.-L. (2009). Extracting discriminative concepts for domain adap-

tation in text mining. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 179–188). Paris, France: ACM.

Daelemans, W., & van den Bosch, A. (2005). *Memory-based language processing*. Studies in Natural Language Processing. Cambridge: Cambridge University Press.

Daelemans, W., Van Den Bosch, A., & Zavrel, J. (1999). Forgetting exceptions is harmful in language learning. *Machine Learning*, *34*, 11–41.

Daumé III, H., & Marcu, D. (2006). Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, *26*, 101–126.

Della Pietra, S., Della Pietra, V., & Lafferty, J. (1997). Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*, 380–393.

Gao, J., Goodman, J., Li, M., & Lee, K.-F. (2002). Toward a unified approach to statistical language modeling for chinese. *Transactions on Asian Language Information Processing*, *1*, 3–33.

Jiang, J., & Zhai, C. (2007). Instance weighting for domain adaptation in NLP. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (pp. 264–271). Prague, Czech Republic: Association for Computational Linguistics.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*, 79–86.

Lee, L. (1999). Measures of distributional similarity. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (pp. 25–32). Maryland, USA: Association for Computational Linguistics.

Lee, L. (2001). On the effectiveness of the skew divergence for statistical language analysis. 8^{th} *International Workshop on Artificial Intelligence and Statistics (AISTATS 2001)* (pp. 65–72). Florida, USA. Online repository http://www.gatsby.ucl.ac.uk/aistats/aistats2001 (Last accessed: March 2013).

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, *37*, 145–151.

Mansour, Y., Mohri, M., & Rostamizadeh, A. (2009). Multiple source adaptation and the Rényi divergence. *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (pp. 367–374). Montreal, Quebec, Canada: AUAI Press.

McClosky, D. (2010). *Any domain parsing: Automatic domain adaptation for natural language parsing*. Doctoral dissertation, Department of Computer Science, Brown University, Rhode Island, USA.

Moore, R. C., & Lewis, W. (2010). Intelligent selection of language model training data. *Proceedings of the ACL 2010 Conference Short Papers* (pp. 220–224). Uppsala, Sweden: Association for Computational Linguistics.

Noreen, E. W. (1989). *Computer-intensive methods for testing hypotheses*. New York, NY, USA: John Wiley.

Pearson, K. (1896). Mathematical contributions to the theory of evolution. – III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London (Series A)*, *187*, 253–318.

Plank, B. (2011). *Domain adaptation for parsing*. Doctoral dissertation, University of Groningen, the Netherlands. Groningen Dissertations in Linguistics 96.

Ravi, S., Knight, K., & Soricut, R. (2008). Automatic prediction of parser accuracy. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 887–896). Honolulu, Hawaii: Association for Computational Linguistics.

Rényi, A. (1961). On measures of information and entropy. *Proceedings of the* 4^{th} *Berkeley Symposium on Mathematics, Statistics and Probability* (pp. 547–561). Berkeley, California, USA: University of California Press.

Sagae, K. (2010). Self-training without reranking for parser domain adaptation and its impact on semantic role labeling. *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing* (pp. 37–44). Uppsala, Sweden: Association for Computational Linguistics.

Van Asch, V. (2012). *Domain similarity measures: On the use of distance metrics in natural language processing*. Doctoral dissertation, University of Antwerp. Available at http://www.clips.ua.ac.be/bibliography/domain-similarity-measures.

Van Asch, V., & Daelemans, W. (2010). Using domain similarity for performance estimation. *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing* (pp. 31–36). Uppsala, Sweden: Association for Computational Linguistics.

van Rijsbergen, C. J. (1975). *Information retrieval*. London, UK: Butterworths.

Verspoor, K., Cohen, K. B., & Hunter, L. (2009). The textual characteristics of traditional and open access scientific journals are similar. *BMC Bioinformatics*, *10*, 1–16.

Zhang, Y., & Wang, R. (2009). Correlating natural language parser performance with statistical measures of the text. *Proceedings of the 32nd annual German conference on Advances in artificial intelligence* (pp. 217–224). Paderborn, Germany: Springer-Verlag.