Year: 2012

Non-negative Matrix Factorisation-based Verb Semantics for 3rd Person Pronoun Resolution

Tuggener, Don ; Klenner, Manfred

Abstract: We present an initial utility study of a distributional model of verb selectional preferences for 3rd person pronoun resolution in German. We investigate cases in which 3rd person pronouns occur as subjects of transitive verbs. In each such case, the likelihood of inserting one of the antecedent candidates is calculated as the conditional probability of the antecedent candidate given either the verb governing the pronoun or the object of the verb. These probabilities are estimated using a matrix derived from frequency counts in a large corpus. Non-negative matrix factorisation is applied as a sort of semantic smoothing to address the sparsity issue inherent in the approach.

DOI: https://doi.org/10.1109/ICSC.2012.64

Originally published at:
Tuggener, Don; Klenner, Manfred (2012). Non-negative Matrix Factorisation-based Verb Semantics for 3rd Person Pronoun Resolution. In: 2012 IEEE Sixth International Conference on Semantic Computing, Palermo, 19 September 2012 - 21 September 2012, 250-254.
DOI: https://doi.org/10.1109/ICSC.2012.64

Non-negative Matrix Factorisation-based Verb Semantics for 3rd Person Pronoun Resolution

Don Tuggener
Institute of Computational Linguistics
University of Zurich
Zurich, Switzerland
Email: tuggener@cl.uzh.ch

Manfred Klenner
Institute of Computational Linguistics
University of Zurich
Zurich, Switzerland
Email: klenner@cl.uzh.ch

***Abstract*—We present an initial utility study of a distributional model of verb selectional preferences for 3rd person pronoun resolution in German. We investigate cases in which 3rd person pronouns occur as subjects of transitive verbs. In each such case, the likelihood of inserting one of the antecedent candidates is calculated as the conditional probability of the antecedent candidate given either the verb governing the pronoun or the object of the verb. These probabilities are estimated using a matrix derived from frequency counts in a large corpus. Non-negative matrix factorisation is applied as a sort of semantic smoothing to address the sparsity issue inherent in the approach.**

***Keywords*-verb semantics; anaphora resolution; distributional semantics;**

I. Introduction

The problem of coreference and pronoun resolution is widely studied as it is an important preprocessing step for higher level NLP applications. Most approaches to coreference resolution rely on pairwise classification of anaphora and antecedent candidates (mention-pair model) or evaluate anaphora on the basis of emerging coreference sets (entity-mention model). Morphosyntactic and semantic features describing the antecedent candidates and the anaphor are collected in feature vectors and processed by rule-based systems or in a machine learning setting. With notable exceptions, selectional preferences of verbs are a lesser studied area in the field. Most approaches in this direction obtain frequency counts from large corpora and, in some cases, apply smoothing to address the sparsity issue inherent in such approaches. However, no approach so far yields substantial improvement by making use of verb selectional preferences.

We present a novel approach to antecedent candidate ranking and selection based on distributional semantics of verb selectional preferences. Opposed to previous work, we encode co-occurrence frequencies of verbs, their subjects, and their objects in a matrix. This allows us to apply non-negative matrix factorisation (NMF) to obtain latent dimensions which we hypothesise model the selectional preferences of the verbs. Multiplying the factorised matrices enables us to smooth the sparse co-occurrence statistics, i.e. obtain non-zero values for unseen verb-noun combinations.

The structure of the paper is as follows. In the next section, we discuss related work which incorporates verb related information in pronoun resolution. We describe the data used in the experiments in section III, followed by an outline of our method and NMF in section IV. Experiments are described in section V and their results are reported in section VI. The paper concludes with a discussion in section VII.

II. Related work

RAPSTAT [1], the first pronoun resolution approach to make use of verb statistics to our knowledge, extended the RAP algorithm [2] by counting how often antecedent candidates occurred with the verb governing the pronoun in a corpus. The antecedent selection by the RAP system was overthrown if a certain threshold was reached. This e.g. allowed the correct resolution of *it* in the sentence: "The Send Message display is shown, allowing you to enter your message and specify where *it* will be sent." RAP had chosen *display* as the antecedent of *it*. RAPSTAT correctly preferred *message* as the antecedent, as it had seen the verb-object pair *send-message* more frequently than the pair *send-display*. [3] implemented RAPSTAT's approach as a postprocessor and as features for a maximum entropy-based pronoun resolution system. They applied Good Turing smoothing to address sparsity. However, the authors did not observe significant improvements over the baseline which did not apply the RAPSTAT extensions. Our approach is similar to [4] which applies latent semantic clustering to frequency counts of subj-verb and verb-object tuples derived from the TuebaD/Z corpus, a German treebank annotated with coreference information [5]. The derived features were appended to a standard feature set for pronoun resolution and evaluated in a machine learning setting [6]. The improvements achieved were marginal. The authors note that sparsity is one of the main issues in such a setting.

Our model borrows from the approach above the idea that frequency counts can be used to model selectional preferences of verbs. It is different in the method, specifically in the way it addresses the sparsity issue.

III. Data

We extracted transitive verbs (based on a frequency threshold of > 10) which have a 3rd person pronoun as their subject from the TuebaD/Z corpus. The entity-mention coreference resolution system presented in [7] generated antecedent candidates (non-pronominal noun phrases) for these pronouns. This yielded 780 verbs with pronouns as their subjects, and a total of 3666 antecedent candidates. We further assured that the true antecedent was among the candidates and we also extracted the direct object. E.g. for the triple "Sie beginnt die Arbeit." (she starts to work), the coreference system generates as antecedent candidates *Strasse, Stadt, Bildhauerin, Herkunft (street, city, sculptress, origin)* from the context, with *sculptress* being the true antecedent and *work* the direct object of the verb *to start*.

As the TuebaD/Z corpus is too small to derive substantial frequency counts, we used the Dewac corpus [8] for this task. We parsed the corpus with the ParZu dependency parser [9], and extracted about 5 million subject-verb-object triples. A quick look at the corpus revealed that 1693 (46.18%) antecedent candidates never occurred with the verbs whose subject position they are deemed fill (e.g. *street* never occurred as the subject of *to start*). Clearly, sometimes the combination of a candidate noun and a verb is unlikely, since the noun violates the selectional restriction of the verb. In other cases, the corpus simply does not feature the subject-verb combination due to the sparsity problem encountered in corpus-based analysis. Class abstraction is needed here, e.g. by relying on the super concept of the noun. In the example, *sculptress* does not occur as the subject of *to start*, but the super concept *artist* does. Such information is available from word nets, but knowledge gaps are to be expected as well. Moreover, ambiguity (e.g. logical metonymy) introduces the need for further decisions and there are no clear-cut decisions in sight. We do not want to discredit approaches based on word nets, we are simply interested in an approach that is completely based on corpus information. Then, of course, we have to cope with sparsity of noun-verb combinations. Latent semantic modelling comes into play, especially approaches that offer a probabilistic interpretation of factorised co-occurrence matrices, namely non-negative matrix factorisation.

IV. Method

Non-negative Matrix Factorisation, NMF [10] is an approach in the field of latent variable modelling that bridges a number of seemingly diverse research directions. It stands in the tradition of Latent Semantic Analysis (LSA), but, since the cell values of the matrix are positive, NMFs can be given a probabilistic interpretation [11]. It has been proven that NMF modelling is equivalent to Probabilistic LSA (PLSA) and k-means clustering [12]. Like LSA, NMFs can be applied to problems that benefit from the analysis of latent semantic dimension. The fact that latent dimensions no longer are forced to be orthogonal is a further advantage over LSA, besides the probabilistic interpretation. Classes no longer need to be strictly exclusive. Dimension reduction is claimed to group objects according to their hidden classes, sometimes called topics. Latent Dirichlet Analysis (LDA) is another technique that strives to detect hidden topics. Consequently, some approaches, e.g. [13], have directly compared NMF to LDA. The task in these systems (and many others from the field of distributional semantics) is related to verb semantics, i.e. the modelling of selectional restrictions with word vectors on the basis of the co-occurrence with other words in a large corpus. Similarity judgements (e.g. based on the cosine between two word vectors) can be used to check for synonymous words or, at the sentence level, for paraphrases. Sophisticated approaches seek to contextualise verb vectors according to the subject or object [14]. Instead of directly comparing single vectors based on a similarity measure, a matrix formed by a fusion of all vectors is used in NMF modelling, e.g. [15].

NMF densifies the vector information with the effect that the original vector dimensions formed by words are transformed into dimensions that now represent word classes. This is the proclaimed effect of these factorisation techniques. In contrast to the work presented here, NMF-based approaches exclusively focus on the benefits of the found latent classes, while we seek to profit from the reconstructed, i.e. the approximated original matrix. Given a factorisation of matrix V in W and H, i.e.

$$V_{n\times m} \approx W_{n\times r} H_{r\times m}$$

V_{approx} is generated by matrix multiplication of $W_{n\times r}$ and $H_{r\times m}$. Depending on the number of iterations, the matrix shape, and other grounds, $V_{approx} \neq V$. NMF minimises an error function, either the Frobenius Norm or the Kullback-Leibler Divergence. The latter is used if language data is modelled, since the former one assumes normal distribution, which is not adequate for natural languages. We are not interested in the latent classes, but in the smoothing effect, i.e. the reduction of zero value cells, they have on the original matrix dimensions. NMF serves as a approximation of the original matrix. If this approximation is based on latent classes, it should act as a kind of semantic smoothing adapting the bare corpus-based frequencies to their hidden inter-dependencies.

V. Experimental setup

We have worked with two matrices. In the first setting, the 780 verbs governing the pronouns form the columns, while the 3666 antecedent candidate nouns form the rows. Each cell counts the number of times the noun occurs as the subject of the verb of the row. This matrix is very

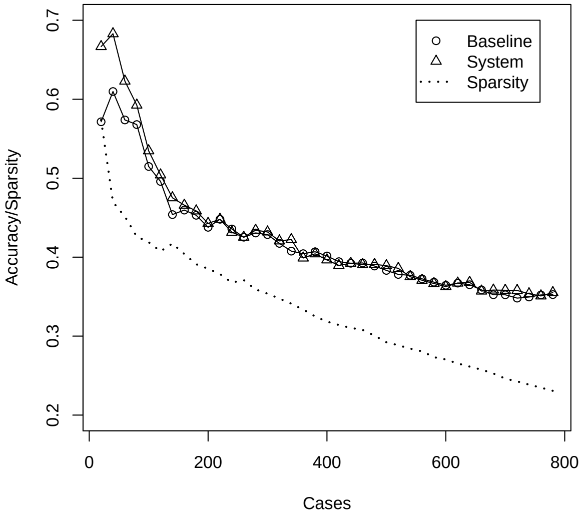

Figure 1. Smoothing the Subject-Verb Matrix

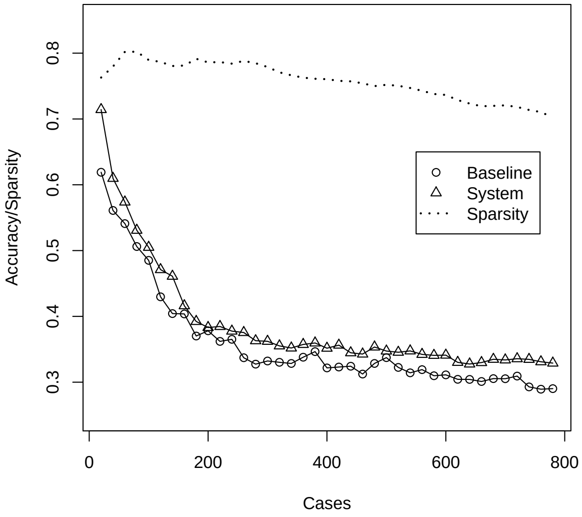

Figure 2. Smoothing the Subject-Object Matrix

sparse. However, sparsity of a noun-verb combination is only fatal where the noun is an antecedent candidate of the verb in question. Since the columns are verbs, factorisation[1] yields verb classes (dimension r in $W_{n\times r}$ and $H_{r\times m}$). Since nouns are subjects of these verbs, the verb classes cluster nouns which are their subjects. Our hypothesis is that this latent semantic class membership of the nouns is preserved when we multiply the factor matrices to obtain the smoothed approximation of the original matrix.

In the second experimental setting, the direct objects form the columns instead of the verbs. Factorisation is supposed to group the subjects in terms of the direct objects they co-occurs with (via the verbs).

As mentioned above, NMF allows a probabilistic view of the matrices. We were interested in conditional probabilities $P(subject|verb)$ and $P(subject|object)$, respectively. We normalised the input matrix V in order to get joint probabilities $P(subject, verb)$ and $P(subject, object)$, and then used a Python implementation of probabilistic non-negative matrix factorisation called “nimfa” [16] to factorise into W and H. Finally, we produced V_{approx} from these matrices and marginalised (column normalisation) in order to get the semantically smoothed conditional probabilities $P(subject|verb)$ and $P(subject|object)$, respectively. In the experiments described in the next section, we selected the best antecedent candidate as the predicted antecedent according to these probabilities.

Since it was clear that the degree of sparsity would influence the performance, we defined a simple preference filter for the test cases that allowed us to reduce sparsity as much as possible in a controlled manner. The filter introduced a score based on noun frequency. For each test case - a verb with its subject candidates - the frequencies of the candidate nouns in the Dewac corpus were summed up. We then processed the cases starting with the best (highest scored) $n = 20$ cases and incremented with $m = 20$ until all of the 780 test cases were reached. Since this method does not provide any information on the overall sparsity of the baseline matrices, we calculated the percentage of non-zero value cells in the baseline matrix as a measure of matrix sparsity.

VI. Results

Fig. 1 visualises the performance of the subject-verb matrix, Fig. 2 the performance of the subject-object matrix. The number of cases is displayed on the x-axis, while accuracy, calculated by the times the system chooses the correct antecedent divided by the number of test cases, is shown along the y-axis. The dotted line depicts matrix sparsity as discussed above. In both curves, accuracy drops as the number of test cases increases. This means when the number of test cases increases, matrix sparsity also increases, and the prediction accuracy decreases. Starting from an accuracy of almost 70% for the $n = 20$ best ranked cases according to our filter, accuracy drops to 30% given all 780 cases. The baseline is defined by the original, unsmoothed matrix. Both the baseline and smoothed matrix produce results significantly better than chance (which is about 20% accuracy, given that we have 4.7 nouns per case). Of course, the crucial questions is whether the smoothed

[1] After some experiments, we found that 5 latent dimensions (rank=5) and 10 iterations (niter=10) worked best.

matrix outperforms the baseline. The answers is yes for the top ranked cases (from $n = 20$ to $n = 100$) in the subject-verb setting. Here, the difference is significant. For instance, given $n = 60$, the baseline accuracy is 57% compared to 62% for the smoothed version. In the subj-obj setting, the smoothed matrix always outperforms the baseline matrix. The accuracy values are generally lower ($<$ 40% after 200 test cases). Matrix sparsity is not problematic in this setting, since the matrix cells are constructed by counting any co-occurrence of two nouns. However, co-occurrence counts become less frequent, i.e. closer to 0 with increasing test cases, which impairs the model. We tried combining the probabilities from experiment 1 and 2 in a third run. The probability of each antecedent candidate was calculated by multiplying the probabilities $P(subject|verb)$ and $P(subject|object)$. This performed worse than the subj-verb setting, though.

We applied a student's t-test, since the differences between the baseline and the factorised version is only obvious for the highest ranked cases. We took the data material from Fig. 1 (comprising 39 accuracy pairs). The p-value is 0.005244, which lets us reject the null hypothesis that the baseline accuracy is greater or equal to the factorised version at the 0.01 significance level.

We might conclude that the smoothed matrix actually better captures the selectional restriction of the verbs than the original matrix. The smoothed matrix does not contain any cells with zero values. We attribute the improvement to this fact. Smoothing fails, on the other hand, if antecedent candidate nouns are a) rare in the sample corpus and b) if the percentage of non-zero value cells in the baseline matrix drops roughly below 40-50%.

Our claim is that word-level sparseness can be overcome by semantic smoothing with the aid of latent modelling, e.g. non-negative matrix factorisation. In our settings, antecedent candidates with zero co-occurrence counts (a kind of out-of vocabulary nouns) for the verb are assigned non-zero values in the smoothed matrix. But does this actually help? Setting 1 and setting 2 seem to support this claim, since a significant, although not very striking, improvement can be observed. In order to analyse the effect more directly, we measured the improvement of exactly those cases in which the zero frequency noun from the baseline matrix turned out to be the most probable antecedent candidate according to the smoothed matrix. Fig. 3 shows how often such a prediction was correct. We see that in most cases an improvement has been achieved. This strongly supports our claim: Matrix smoothing is able to compensate word-level sparsity. The question, why the difference to the baseline is that low for $n > 200$ (cf. Fig.1), remains. The answer is that smoothing also affects baseline matrix values > 0 and here the positive effect seems to get absorbed to a certain degree. A baseline matrix with less than roughly 40-50% non-zero value cells seems not to be a solid enough ground for the NMF approach to be effective in our setting. Additional methods of reducing the number of zero value cells in the co-occurrence matrix are needed.

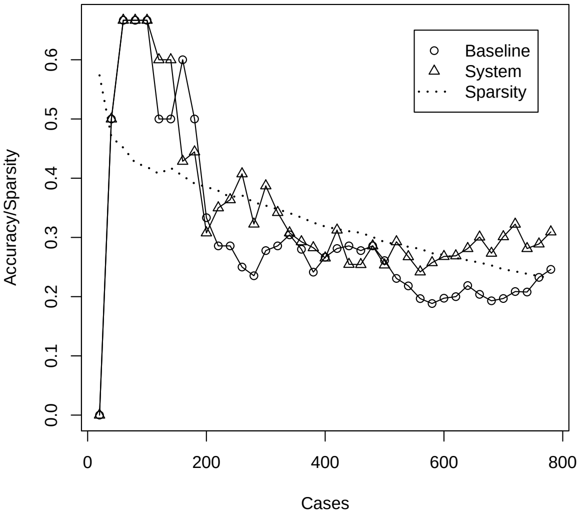

Figure 3. Effect of Smoothing on Out-of Vocabulary Nouns

VII. Conclusion

We have introduced an approach to model selectional restrictions of verbs with non-negative matrix factorisation. Estimation methods for selectional restrictions that solely rely on frequency data run into problems caused by data sparsity, as one cannot expect to find all permissible filler objects of a verb, even in a large corpus. This is especially problematic in languages featuring (sometimes rare) compound nouns, such as German. Class abstraction based on word nets or decomposition might alleviate the problem to a certain degree. What is generally needed is a kind of semantic smoothing. That is, a method that clusters the known filler objects and utilises the resulting latent dimensions to redistribute class membership weights (e.g. in form of conditional probabilities).

We aim to utilise our model in a coreference resolution system, in which the antecedent nouns for a personal pronoun must be licensed by the verbal head of the pronoun. Our experiments show that good results are to be expected if the input matrix is not too sparse. Note that we now talk about *matrix sparsity* and not about *sparsity of single words*. Word sparsity, this is our hypothesis, can be overcome if non-sparse matrices can be constructed. We have presented some evidence for this hypothesis. Future work is devoted to elaborate on these initial findings.

References

[1] I. Dagan, J. Justeson, S. Lappin, H. Leass, and A. Ribak, "Syntax and lexical statistics in anaphora resolution," *Applied Artificial Intelligence*, vol. 9, pp. 633–644, 1995.

[2] S. Lappin and H. J. Leass, "An algorithm for pronominal anaphora resolution," *Computational Linguistics*, vol. 20, pp. 535–561, 1994.

[3] A. Kehler, D. Appelt, L. Taylor, and A. Simma, "The (non)utility of predicate-argument frequencies for pronoun interpretation," in *ACL*, 2004, pp. 289–296.

[4] E. W. Hinrichs and H. Wunsch, "Selectional preferences for anaphora resolution," in *Theory and Evidence in Semantics*, E. Hinrichs and J. Nerbonne, Eds. CSLI Publications, 2009.

[5] H. Telljohann, E. Hinrichs, and S. Kübler, "The tüba-d/z treebank - annotating german with a context-free backbone," in *LREC*, 2004, pp. 2229–2235.

[6] H. Wunsch, "Rule-based and memory-based pronoun resolution for german: A comparison and assessment of data sources," Ph.D. dissertation, Universität Tübingen, 2010.

[7] M. Klenner and D. Tuggener, "An incremental entity-mention model for coreference resolution with restrictive antecedent accessibility," in *RANLP*, 2011, pp. 178–185.

[8] M. Baroni and A. Kilgarriff, "Large linguistically-processed web corpora for multiple languages," in *EACL*, 2006, pp. 87–90.

[9] R. Sennrich, G. Schneider, M. Volk, and M. Warin, "A New Hybrid Dependency Parser for German," in *Proc. of the German Society for Computational Linguistics and Language Technology 2009 (GSCL 2009)*, Potsdam, Germany, 2009, pp. 115–124.

[10] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *NIPS*, 2000, pp. 556–562.

[11] M. Shashanka, B. Raj, and P. Smaragdis, "Probabilistic latent variable models as nonnegative factorizations," *Computational Intelligence Neuroscience*, 2008.

[12] É. Gaussier and C. Goutte, "Relation between plsa and nmf and implications," in *SIGIR*, 2005, pp. 601–602.

[13] G. Dinu and M. Lapata, "Measuring distributional similarity in context," in *EMNLP*, 2010, pp. 1162–1172.

[14] S. Thater, H. Fürstenau, and M. Pinkal, "Contextualizing semantic representations using syntactically enriched vector models," in *ACL*, 2010, pp. 948–957.

[15] T. Van de Cruys, T. Poibeau, and A. Korhonen, "Latent vector weighting for word meaning in context," in *EMNLP*, 2011, pp. 1012–1022.

[16] M. Zitnik and B. Zupan, "Nimfa: A python library for nonnegative matrix factorization," *Journal of Machine Learning Research*, vol. 13, pp. 849–853, 2012.