Which Factors Contributes to Resolving Coreference Chains with Bayesian Networks?

Davy Weissenbacher, Yutaka Sasaki

Which Factors Contributes to Resolving Coreference Chains with Bayesian Networks?

Davy Weissenbacher[1] Yutaka Sasaki[2]

[1] IRISA (INRIA, University of Rennes 2, INSA, CNRS) - Rennes, France
davy.weissenbacher@gmail.com
[2] Toyota Technological Institute-CoIN Laboratory
2-12-1 Hisakata, Tempaku, Nagoya, 468-8511, JAPAN
yutaka.sasaki@toyota-ti.ac.jp

Abstract. This paper describes coreference chain resolution with Bayesian Networks. Several factors in the resolution of coreference chains may greatly affect the final performance. If the choice of machine learning algorithm and the features the learner relies on are largely addressed by the community, others factors implicated in the resolution, such as noisy features, anaphoricity resolution or the search windows, have been less studied, and their importance remains unclear. In this article, we describe a mention-pair resolver using Bayesian Networks, targeting coreference resolution in discharge summaries. We present a study of the contributions of comprehensive factors involved in the resolution using the 2011 i2b2/VA challenge data set. The results of our study indicate that, besides the use of noisy features for the resolution, anaphoricity resolution has the biggest effect on the coreference chain resolution performance.

Key words: Coreference Resolution, Anaphoricity Resolution, Bayesian Networks, Clinical Informatics

1 Introduction

Anaphora is a linguistic relation between two textual entities, which are commonly named mentions. The relation is defined when an entity, *the anaphor*, refers to another one, *the antecedent*. For example, in the sentences

[Mr. TTT]$_1$ *was brought to [the operating room]*$_2$ *where [he]*$_1$ *underwent [a coronary artery bypass graft]*$_3$ *x 3. [The patient]*$_1$ *tolerated [the procedure]*$_3$ *well.*

the pronoun *[he]*$_1$ is the anaphor, and it refers to the Noun Phrase (NP), *[Mr. TTT]*$_1$. When both mentions of the anaphoric relation refer to an identical object of the world, the relation is said to be coreference. As coreference is an equivalence relation, all mentions can be partitioned into different classes called *coreference chains*. In our example we have two coreference chains subscripted 1 and 3. The NP {*the operating room*} is a singleton and does not form a chain.

The resolution of coreference chains is still a difficult task. Whereas several factors are co-dependent in the resolution and may greatly affect the final performance when not set up correctly, only a few of them received specific attention

in previous studies. While (1) the choice of the Machine Learning (ML) framework and (2) the features the ML algorithm relies on are largely addressed by the community, (3) the impact of the noise of the features, (4) the quality of the anaphoricity resolution and (5) the optimal size of the search windows, which are crucial in the mention-pair resolution strategy, have been less studied and their respective impacts on the resolution remain unclear.

The Informatics for Integrating Biology and Bedside (i2b2) institute has been holding a series of annual challenges to compare NLP systems on various tasks in the medical domain. The fifth i2b2/VA challenge, held in 2011, was on coreference resolution. While designing our own resolution system, we proceed to a comprehensive study of the effects of the above five factors on the overall performance of our system. The main contributions of this article are (1) to describe a mention-pair resolver based on a Bayesian Network addressing coreference resolution in discharge summaries and (2) to evaluate the direct effect of each factor on the overall resolution to guide further research by giving the highest priority to the most effective one.

In the following Section 2, we describe the resolver implemented and the features driving the classification. The corpus, the metrics and the protocol used for the experiments are detailed in Section 3. Impacts of the factors are discussed in Section 4. Section 5 presents related work, and the last section concludes our study.

2 Resolving Coreference Chains

2.1 Preprocessing

To preprocess the i2b2/VA corpus, we use an annotation platform integrating publicly available annotation modules. It recognises the logical structures, *i.e.* titles, paragraphs, etc., thanks to handmade Regular Expressions (REs). As the sentence segmentation is crucial for anaphora resolution we used the preformatted sentences provided by the challenge organizers. To segment the words and produce a shallow parsing analysis of the documents (POS tagging and Chunking), we have chosen the Genia Tagger[3]. The pre-annotated concepts in the i2b2 corpora can be thought similar to Named Entities, we relied entirely on those concepts. The syntactic analysis of the sentences and the grammatical roles have been extracted by *Enju*[4]. Heads of NPs also play an important role in resolution since lots of features are computed based on them. To ensure good precision, NP and VP chunks are submitted and analysed separately from the whole sentence by *Enju*[5]. When the chunk analysis fails, heuristics are used [1]. Many resources have been developed for the Medical domain, we applied MetaMap[6] to automatically extract concepts of this domain.

[3] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

[4] http://www-tsujii.is.s.u-tokyo.ac.jp/enju/

[5] Extracting heads from the analysis of the full sentence gave bad results during preliminary experiments.

[6] http://metamap.nlm.nih.gov/

2.2 Resolution Strategy

In a traditional approach to resolve coreference chains, two steps can be distinguished, the anaphoricity resolution followed by the coreference resolution.

Anaphoricity resolution consists of distinguishing anaphoric phrases which expect an antecedent from other phrases for which any suggestion of an antecedent would result in an error. Non-anaphoric phrases are, for example, pleonastic phrases (*e.g. It would be fine to...* vs *I have reviewed it...*), deictic phrases (*e.g.*, in our corpus, *this report, this year*) or the first NPs in coreference chains (first mentions of an object referred to by a chain are not anaphoric by definition).

The coreference resolution aims to build the coreference chains; all mentions referring to the same object should be included in a unique chain. When a strategy based on clustering is not chosen, a strategy relying on a binary classification is possible. For each anaphoric mention, considered in order, a list of previous mentions occurring in a search window is created, and one candidate in the list is chosen as antecedent. In the usual model, the *mention-pair* model, only pairs composed of an anaphoric mention and its respective candidates are described. Each pair received a score, and the candidate of the best pair is taken as antecedent. Once all pairs have been resolved, chains are built during an additional process, usually by taking the transitive closure with respect to the semantic constraints within the chains.

Classification methods are easy to use with the mention-pair model. We chose this model for our system. To build the coreference chains we took the transitive closure of the coreferent pairs. Incoherences within the resulting chains are post-edited by taking in the list of scored pairs the candidate of the first pair which agrees with the semantic constraints of the chain.

2.3 Features and Classifiers

In our system, pairs of mentions are described with a set of 32 features. They are features commonly used for coreference resolution plus features specific to the genre of our documents.

Our features can be separated into 3 categories: lexical, syntactic and semantic. Lexical features aggregate information about number, gender, position and all matching based features (string matching, embedded NPs, repeated NP etc.). Syntactic features provide information about grammatical roles of the mentions, syntactic parallelism or collocation patterns. Ground truth mentions annotated in the corpus are classified into 5 types of concepts: *person, problem, treatment, test, other*. From these semantic annotations we acquire reliable features and express constraints of coherence. Among the mentions denoting persons we specify, using handmade REs, the main protagonists of the discharge, namely the patient, his/her family, doctor and medical services. Mentions which do not refer to people are described in greater detail based on the MetaMap categories they match.

Pronouns have separate resolution procedures as they carry different information than NP mentions and tend to resolve with the closer candidates. We make use of 23 of the previous features to model the salience of the candidates as described in [1], except for the pronouns "I" and "we" which, in our corpus, are likely to resolve with the closest mention of the doctor.

To carry out the classification we select the Bayesian Network (BN) framework, a Machine Learning framework adapted to the distinctive characteristics of Natural Language Processing (NLP) tasks [2]. A BN is a probabilistic graphical model. It is composed of a qualitative description of the dependencies between a set of random variables, represented by an oriented acyclic graph, and of a quantitative description, a set of Conditional Probability Tables (CPT) where each random variable is associated to a graph node. For each of the previous features a random variable is created and the conditional probability table associated to the random variables gives information about which features it influences and is influenced by. In all our experiments the structure of the graph and the values of the CPTs are automatically learned from the data.

Because a coreference relation is an equivalence relation, positive and negative examples submitted to the machine learner during induction have to be carefully selected [3, 4]. Positive examples are anaphoric mentions linked to their closest immediate mention which belong in the same coreferent chain. A negative example is a pair of 2 mentions belonging to 2 different chains. We removed the trivial negative examples and presented only the 3 best negative examples for each anaphor to the system. The best negative examples are obtained during a preprocessing stage. The BN is trained iteratively 3 times, using the best pairs of the previous iteration.

As a first working hypothesis, our BN has been trained using the score-based algorithm K2 with a local metric, limited to 5 parents without imposing the Naive Bayes structure, combined with a *maximum a posteriori estimation*, the alpha parameter set to 0.5.

3 Experiments

3.1 Corpus and Metrics

The corpus released for the 2011 i2b2/VA challenge is composed of discharge summaries provided by four health institutes. We worked with a subpart of the corpus, 251 documents for training and 173 for testing, referred to by the organizers as the i2b2/UPMC in [5].

To evaluate our system and compare it with other participants' systems, we used the official evaluation tool. By comparing the chains proposed by our system with the gold standard, this tool calculates 3 different metrics and their unweighted average: B^3, MUC, and CEAF. A presentation of those metrics and a discussion about their respective deficiencies can be found in [5].

3.2 Protocol

When the mention-pair strategy is applied, five factors can directly influence the performance of the coreference resolution. The first factor is the choice of the features to describe mentions and pairs. When the features are computed automatically, the second factor is the noise of the feature values. This type of noise can strongly degrade the induction process. As shown in [1] it might be better to do without a feature if it cannot be computed above a certain threshold of accuracy. Once the most reliable features have been selected, the machine learning framework, the third factor, has to be accurately chosen to ensure a good compromise between the power of expression required to learn the rules and the search for the optimal solution in the corresponding hypothesis space. The fourth factor is the choice of a strategy for resolving the anaphoricity. Whereas the anaphoricity resolution and the coreference resolution are co-dependent, the former has only lately received interest from the community [3]. The last factor is the size of the search window. It determines which mentions should be inserted in the list of possible candidates; the optimal size can never be known in advance since it depends on the genre and the domain of the corpus.

To optimise our coreference resolver we run a set of experiments, changing the value of one factor at a time, in order to find the more effective factors for the resolution. The next section presents the results obtained for each factor.

4 Evaluation

Noisy Features Factor. The impact of noisy features in resolution has been studied during the i2b2/VA challenge with a special track (track 1A). This track evaluates end-to-end resolution systems. With ground truth mentions hidden from the resolvers, a drop in performance of the systems ranging from 10.3% to 39.0% has been observed [5]. Noisy features appear to be the most critical factor to perfect in order to achieve a suitable score in coreference resolution. We did not carry out additional experiments for this factor.

Machine Learning Framework Factor. With the current progress of Machine Learning, many frameworks are now available, making the choice of a particular framework difficult. Their advantages depend on the number, the type and the structure of the data points the induction has to be run on. In the experiment below we intended to estimate the scale of the gain (or loss) by changing from one framework to another (even if some limitations have to be supported to use a particular framework). We have selected 4 frameworks broadly used in NLP: a decision tree classifier, a SVM classifier, a Naive Bayes and Bayesian Network classifiers.

For this experiment, all systems have their factors parametrised identically except for the classifiers they rely on to score the pairs of mentions[7]. All anaphoric

[7] We use the Weka machine learning tools. Each machine learning framework can be tuned to improve the induction, but we used the default options, except for the Bayesian Network where the default option is the Naive Bayes structure.

mentions are given to the coreference resolver. The windows size is the largest possible, all anaphoric mentions which occur before the anaphoric mention to resolve are considered. To estimate the improvement, we report the performance of the baseline resolver published in [5]. The baseline resolver predicts all mentions as singletons.

Table 1 is quite revealing in two ways. First, it shows that there is a benefit of using an adapted ML framework. If all ML frameworks outperform the baseline system, there is a big difference in performance between the SVM and the BN, 7.8% in F-measure. The features used to model NLP data are strongly dependent due to the nature of Natural Language itself. The BN is the only classifier able to represent those dependencies and consequently makes a better discrimination between the mentions. The Naive Bayes classifier, helped by its knowledge of the prior probability of the features, is less sensitive to the missing values which are frequent with the features used for coreference resolution (*e.g.* unknown gender or grammatical roles). The default polynomial kernel support vector machine classifier proposed in Weka for the SVM classifier gives deceiving results, unusually worse than the decision tree. Better parameters or dedicated kernels should allow better results.

Secondly, the score of the BN, F=0.921, is higher than the score of the best system of the i2b2/VA challenge F=0.913 (P=0.905, R=0.920)[6], whereas our system does not make as extensive use of domain dependent knowledge as the latter system does. This fact supports the conclusion in [7] and is important since it demonstrates that an acceptable score can be achieved on this corpus using domain independent knowledge. However this result is only possible if the anaphoric mentions are perfectly resolved by the system. This perfect resolution is for the moment out of reach, even though resolving anaphoricity is much easier than resolving coreference [8], see section 4 for further discussion.

Systems	P	R	F
Baseline	.523	.602	.548
Decision Tree	.859	.850	.854
SVM	.849	.839	.843
Naive-Bayes	.894	.912	.903
Bayesian Network	**.912**	**.930**	**.921**

Table 1. Coreference resolution on Test corpus with various machine learning frameworks (Anaphoric mentions are revealed, search window set to all previous candidates)

Feature Selection Factor. Features are central for the resolution because they express constraints/preferences to choose/discard a mention as antecedent. Therefore, they are the main subject of study for the coreference resolution. According to Zheng and *al.* [3] their number may largely vary within the range

from 8 to 134. Their nature also is still under discussion: domain dependent features *vs.* general features.

In this study we give preference to domain independent features supplemented by semantic features adapted to the specific genre of our documents. The discharge summaries follow a specific scenario. A main actor, the patient, interacts with a few other characters, Doctor or medical services for instance, and whose body is described in detail. This causes a predominant chain of coreference, the chain of the patient, and several short coreferent chains. As for other participants' systems, our system relies on the categories associated with the mentions and tries to refine those categories. For the person category we wrote REs to discriminate the patient from the doctor, the family and medical services (*Coherent Roles* features in Table 2). To refine other categories we use the best UMLS concepts assigned by the word sense disambiguation module of the MetaMap tool (*Coherent Medical Concepts* features). Finally, we use the likelihood computed on the training corpus for two heads of mentions to be coreferent (*Heads Coreferent Mentions* features). Like Rink and *al.* [7], we believe this strategy can be applied to all documents with similar scenarios (accident reports or encyclopedia articles, are possible examples).

The ablation study in Table 2 confirms the contribution of each feature. It suggests that the set of features added does not induce an important improvement of the overall score, only 2.8%. A similar conclusion can be drawn from Xu and *al.*'s [6] experiment. The best score is achieved by the lexical feature based system. Adding syntactic features does not improve the resolution and may even degrade the performance[8]. Semantic features improve the recall, particularly of medical concepts, but it is at the cost of a lower precision.

Bayesian Network	P	R	F
Lexical Feature	.927	.927	.927
Syntactic Feature			
+ Grammatical Roles	.910	.907	.909
+ Syntactic Parallelism	.910	.907	.909
+ Simple Collocations	.905	.903	.904
+ Syntactic Collocations	.902	.903	.902
Semantic Feature			
+ Coherent NEs	.902	.899	.900
+ Coherent Roles	.907	.905	.906
+ Coherent Medical Concepts	.912	.929	.921
+ Heads Coreferent Mentions	.912	.930	.921

Table 2. Ablation study on the features used by the BN, performed on the Test corpus (Anaphoric mentions are revealed, search window set to all previous candidates)

[8] However syntactic features seem to corroborate the semantic ones. When our BN exploits lexical and semantic features without the syntactic ones, it performs worse than when it exploits all features, F=0.901 against F=0.921, respectively.

Anaphoricity Accuracy Factor. The good performance of our system is mainly due to the perfect anaphoricity resolution. To calculate its impact on the coreference resolution we introduced noise in the anaphoricity resolution. The anaphoricity resolver decides if a particular mention admits an antecedent or not; it does not have to find which mention is the antecedent. The quality of this resolution is crucial. False positives are mentions resolved as anaphoric when they are not and cause the coreference resolver to create new chains or include the false positives in any existing chain. False negatives are anaphoric mentions not recognized as such by the resolver and may result in a drop of recall if these anaphoric mentions are not chosen as antecedents for other anaphoric mentions.

The current state-of-the-art scores range around 80% accuracy (Acc.) for a general domain corpus [4]. In order to evaluate the easiness of the task on our corpus, we have implemented a baseline anaphoricity resolver. Due to space limitations, we will not describe the anaphoricity resolver in detail. This resolver is also based on a Bayesian Network and performs two different resolutions for Definite NPs and for pleonastic pronouns *it, this, that, what, which* (other pronouns in our resolution are always considered anaphoric).

To classify a given Definite NP, features used are targeting possible synonyms which occur before the NP in the document. The synonyms are found based on string matching, edit distance, the WordNet dictionary, the MetaMap concepts of both mentions and the sections where possible synonyms appear. Sections are important in the discharge summaries since they indicate how to interpret following paragraphs, a context which is mandatory to resolve some coreferences. This can be illustrated briefly by two occurrences of *CVA* appearing in the section *History of Present Illness* and the section *Family History*; they are synonyms but they cannot be coreferent. Pleonastic pronouns *it* and *this* are detected by the filter described in [1] and adapted for our corpus. Other pleonastic pronouns are classified according to their immediate context. A pronoun, like the pronoun *what*, when immediately preceded by a noun tends to be anaphoric, whereas preceded by a verb is more likely to be non-anaphoric.

Despite its simplicity our anaphoricity resolver reaches a decent score of 87.6% Acc. on the test corpus. Preliminary investigation of the results shows that the number of false negatives is much higher than the number of false positives, 2516 against 881. This is mainly due to the lack of the resources which are needed to establish synonym links between acronyms (such as *transesophageal echocardiogram* and *TEE*), hyperonyms (*examination* and *endoscopy*) or drug names (*lipitor* and *Atorvastatin*). General resources like Wikipedia has been found valuable resources [9, 7, 6] to provide such knowledge to the resolver.

Table 3 presents the coreference resolution achieved with varying quality of anaphoricity resolutions. According to predefined thresholds, we have corrupted gold anaphoric mentions to non-anaphoric and vice-versa. Mentions have been chosen randomly except for those which are preceded by a mention which exactly matches or has a similar head. Last constraints hold to avoid to corrupt anaphoric mentions which can be detected with a high precision. Bold scores are the score obtained when using the outputs of our anaphoricity resolver.

From the data in Table 3 it is apparent that the biggest improvement is made by ameliorating the anaphoricity resolution with a possible gain of 12.7% in F-measure. Given the current performance of our anaphoricity resolver, 13.4% error rate, our coreference resolver reaches the top performance achieved during the last i2b2/VA challenge, with a score which is about equal to the score of the 9^{th} system of the competition (a total of 20 teams participated in).

Surprisingly, our system obtains a similar score when the noise threshold of is set to 10%. A possible explanation for this might be that in our experiment errors are randomly distributed, regardless of the easiness of the anaphoricity resolution. Whereas mentions incorrectly classified by our anaphoricity resolver are often the most difficult mentions to assign in chains.

BN Performances	P	R	F
noise level			
0%	.912	.930	.921
5%	.913	.877	.895
10%	.892	.828	.857
13.4%	**.829**	**.891**	**.857**
15%	.881	.784	.826
20%	.862	.746	.794

Table 3. Coreference resolution performances on the Test corpus for the BN given various anaphoricity resolutions *(in Accuracy)*

Search Window Factor. The last factor is the size of the search window. The bigger the size of the window is, the higher is the risk to choose a "better" candidate, that is, a candidate different from the antecedent. While if the window is too small, none of the coreferent mentions may be found in the list of the candidates. The optimal size depends on the genre and the domain of the corpus [10]. In the discharge summaries, a list of medications or medical history report may separate an anaphor from its coreferent mentions by hundreds of sentences. The highest distance found in the training corpus was 274 sentences.

We have computed the search window as a percentage of sentences which have to be explored before finding the closest coreferent mention of each anaphoric mention. The ratios of antecedents captured by the search windows have been computed directly on the test corpus[9]. Supplementary analysis shows that 20.3% are intrasentential anaphora in the test corpus (*resp.* 22.8% in the training corpus), 50.4% of the antecedent are located in the previous sentence (*resp.* 54.3%) and, as suggested by Zheng and *al.* [3], if the window is fixed as usual to the 10

[9] Similar computations on the training corpus have been done and show a difference of 7%. That is, a window of 83% on the training corpus is enough to capture all antecedents.

previous sentences only 76.3% (*resp.* 79.2%) of the antecedents could have been found.

Table 4 summarizes the performance of the coreference resolver according to various sizes of windows. It appears that optimizing the size of the search window improved the performance of the resolution. Whereas the recall of the system sees no change, the precision, in reducing the number of candidates, has a consequent rise of 1.6%. This leads to the overall improvement of the system which does slightly better than the lexical based resolver described in Section 4.

However examining such proportion of document is still not satisfying. Many algorithms, for example based on centering[11] or the attention of the reader [12], have been proposed to update dynamically the list of candidates by removing from it impossible or old candidates. To test the interest of such algorithms we run a last experiment. We fixed a smaller size for the search window, set to the 10 previous sentences, and we artificially introduced the last coreferent mention. This experiment evaluates the capacity of the resolver to choose the coreferent mention among a few candidates and it suggests maximum scores reachable for the coreference resolution with our current features. With this last configuration the system's score reaches F=0.931.

BN performances with different search windows		P	R	F
Window size	*Antecedents captured*			
41%	94.55%	.906	.908	.907
67%	99.04%	.925	.926	.925
73%	99.62%	.926	.926	.926
90%	100%	.928	.929	.929
10 sentences with antecedents appended		.918	.934	.931

Table 4. Coreference resolution performance on the Test corpus for the BN given various search window sizes

5 Related Work

Our system is inspired from earlier modular strategies for resolution proposed by Rich and LuperFoy [13] or Mitkov [10]. Our approach targeting the patient and specialising other mention types is close to the general approach taken by the competing systems during the i2b2/VA Challenge [5]. Many of our features are similar to those described in [14].

Effects on the coreference resolution of several factors discussed in this article have been the main focus of several existing studies. While [15] examines possible discriminant features for clinical documents, the choice of features is still a significant problem for coreference resolution [16], [10] tests the benefits of using heuristics when the features are not available. Induction performed through various ML frameworks is studied by [17] for supervised methods. Advantages

of sophisticated models compared to pairwise model resolution are criticized by Bengtson and Roth [4]. Finally, during their study to predict the difficulty of the coreference resolution on corpora, Stoyanov and *al.* [18] investigate possible performance improvements allowed by a better anaphoricity resolution and a better detection of the mentions. However, those studies often made comparisons between systems which differ by several factors at a time. In his extensive study about anaphora resolution, [10] draws our attention to the difficulties for making direct comparison between two coreference resolvers. If the systems are usually working on the same corpus, the preprocessing and the implementation of the features, for example, are rarely similar and introduce bias in the comparison. We are not aware of any existing study which carries out an exhaustive enquiry on the role of each factor for a given resolver. This article is an attempt to clearly measure the influences of the most important factors in the resolution.

Conclusion

In this article we introduced a promising coreference resolver based on a Bayesian Network and we presented a comprehensive study of the contribution of all important factors involved in the resolution.

Our system, to resolve coreference relations in clinical documents, relies on the mention-pair resolution strategy and uses a Bayesian Network to score the anaphoric pairs. The set of features implemented are features commonly used by ML based systems, completed with semantic features specialized for the genre of our documents. The semantic features track down the main objects of the discourse and express constraints by specifying the concepts these objects belong to. Using a basic anaphoricity resolver, we achieved an F-score of 0.857 on the 2011 i2b2/VA Challenge data set on coreference resolution.

By investigating the factors that contribute to the coreference resolution, our intention was to give a precise evaluation of their individual contributions to overall performance. Besides the use of noisy features for resolution, anaphoricity resolution has the biggest effect on the performance since both resolutions are strongly co-dependent. The choice of the ML framework can also strongly affect the results. The genre of the documents necessitate to adapt the size of the search window. Finally, the choice of the features, while main interest of the community, appears to be the less important factor in term of possible gain for resolution.

These findings suggest several courses of action for further enhancement of our resolver, with first priority given to our anaphoricity resolver. Based on Wikipedia, we are currently studying analogy distances between two mentions. By capturing valuable synonym relations this addition not only may largely improve our anaphoricity resolver, but also the coreference resolver. In the short-term the BN used for the anaphoricity resolution will be merged with the BN used for the coreference resolution in order to determine jointly both resolutions [8]. At medium term we will make use of Bayesian Logic Programs capable of representing all mentions and their associated chains within a unique probabilis-

tic model, abolishing thus the unjustified independence assumption between the candidates, an assumption imposed by the current BN framework.

Acknowledgments

We thank Pr. Gina-Anne Levow for her helpful comments and remarks.

References

1. Weissenbacher, D., Nazarenko, A.: Comprendre les effets des erreurs d'annotations des plates-formes de tal. Traitement Automatique des Langues **52** (2011) 161–185
2. Behera, L., Goyal, P., McGinnity, T.: Application of Bayesian Framework in Natural Language Understanding. IETE Technical Review **25** (2008) 251–269
3. Zheng, J., Chapman, W., Crowley, R., Savova, G.: Coreference resolution: A review of general methodologies and applications in the clinical domain. Journal of Biomedical Informatics **44** (2011) 1113–1122
4. Bengtson, E., Roth, D.: Understanding the value of features for coreference resolution. In: EMNLP'08. (2008)
5. Uzuner, O., Bodnari, A., Shen, S., Forbush, T., Pestian, J., South, B.: Evaluating the state of the art in coreference resolution for electronic medical records. Journal of the American Medical Informatics Association (2011)
6. Xu, Y., Liu, J., Wu, J., Wang, Y., Tu, Z., Sun, J., Tsujii, J., Chang, E.I.C.: A classification approach to coreference in discharge summaries: 2011 i2b2 challenge. Journal of the American Medical Informatics Association (2012)
7. Rink, B., Roberts, K., Harabagiu, S.: A supervised framework for resolving coreference in clinical records. Journal of the American Medical Informatics Association (2012)
8. Denis, P., Baldridge, J.: Joint determination of anaphoricity and coreference resolution using integer programming. In: Proceedings of NAACL. (2007) 236–243
9. Gooch, P., Roudsari, A.: Lexical patterns, features and knowledge resources for coreference resolution in clinical notes. Journal of Biomedical Informatics (2012)
10. Mitkov, R.: Anaphora Resolution. Longman(Pearson Education) (2002)
11. Grosz, B., Weinstein, S., Joshi, A.: Centering: a framework for modeling the local coherence of discourse. Computational Linguistics **21** (1995) 203–225
12. Strube, M.: Never look back: An alternative to centering. In: 17th International Conference on Computational Linguistics. Volume 2. (1998) 1251–1257
13. Rich, E., LuperFoy, S.: An architecture for anaphora resolution. In: Proceedings of the second conference on Applied natural language processing. (1988) 18–24
14. Zweigenbaum, P., Wisniewski, G., Dinarelli, M., Grouin, C., rosset, S.: Résolution des coréférences dans des comptes rendus cliniques. Une expérimentation issue du défi i2b2/VA 2011. In: Actes de RFIA. (2012)
15. He, T.: Coreference resolution on entities and events for hospital discharge summaries. Master's thesis, MIT (2007)
16. Preiss, J.: Machine learning for anaphora resolution. (2001)
17. Ng, V.: Supervised noun phrase coreference research: The first fifteen years. In: 48th Annual Meeting of the ACL. (2010) 1396–1411
18. Stoyanov, V., Gilbert, N., Cardie, C., Riloff, E.: Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. In: Proceedings of the 47th Annual Meeting of the ACL. (2009) 656–664