
Abstract

This thesis presents a machine learning approach to the resolution of coreferential relations between nominal constituents in Dutch. It is the first automatic resolution approach proposed for this language. The corpus-based strategy was enabled by the annotation of a substantial corpus (ca. 12,500 noun phrases) of Dutch news magazine text with coreferential links for pronominal, proper noun and common noun coreferences. Based on the hypothesis that different types of information sources contribute to a correct resolution of different types of coreferential links, we propose a modular approach in which a separate module is trained per NP type. Lacking comparative results for Dutch, we also perform all experiments for the English MUC-6 and MUC-7 data sets, which are widely used for evaluation.

Applied to the task at hand, we focus on the methodological issues which arise when performing a machine learning of language experiment. In order to determine the effect of algorithm ‘bias’ on learning coreference resolution, we evaluate the performance of two learning approaches which provide extremes of the eagerness dimension, namely TIMBL as an instance of lazy learning and RIPPER as an instance of eager learning. We show that apart from the algorithm bias, many other factors potentially play a role in the outcome of a comparative machine learning experiment. In this thesis, we study the effect of selection of information sources, parameter optimization and the effect of sampling to cope with the skewed class distribution in the data. In addition, we investigate the interaction of these factors.

In a set of feature selection experiments using backward elimination and bidirectional hillclimbing, we show the large effect feature selection can have on classifier performance. We also observe that the feature selection considered to be optimal for one learner cannot be generalized to the other learner. Furthermore, in the parameter optimization or model selection experiments, we observe that the performance differences within one learning method are much larger than the method-comparing performance differences. A similar observation is made in the experiments exploring the interaction between feature selection and parameter optimization, using a genetic algorithm as a computationally feasible way to achieve this type of costly optimization. These experiments also show that the parameter settings and information sources which are selected after optimization cannot be generalized. In the experiments varying the class distribution of the training data, we show that both learning approaches behave quite differently in case of skewedness of the classes and that they also react differently to a change in class distribution. A change of class distribution is primarily beneficial for RIPPER. However, we observe that once again no particular class distribution is optimal for all data sets, which makes this resampling also subject to optimization.

In all optimization experiments, we show that changing any of the architectural variables can have great effects on the performance of a machine learning method, making questionable conclusions in the literature based on the exploration of only a few points in the space of possible experiments for the algorithms to be compared. We show that there is a high risk that other areas in the experimental search space lead to radically different results and conclusions.

At the end of the thesis, we move away from the instance level and concentrate on the coreferential chains reporting results on the Dutch and English data sets. In order to gain an insight into the errors committed in the resolution, we perform a qualitative error analysis on a selection of English and Dutch texts.

Samenvatting

Dit proefschrift gaat over het gebruik van lerende technieken voor de resolutie van coreferentiële relaties tussen nominale constituenten in het Nederlands. Het is meteen de eerste automatische aanpak voor deze taal. Die corpusgebaseerde aanpak werd mogelijk gemaakt door de annotatie van een aanzienlijk corpus van teksten uit een Vlaams weekblad met nieuws uit de nationale en internationale actualiteit. Tijdens de annotatie werden ongeveer 12,500 nominale constituenten, bestaande uit eigennamen, soortnamen en pronomina, voorzien van coreferentiële informatie. Uitgaande van de hypothese dat het type informatie dat nodig is voor een correcte resolutie kan verschillen per type coreferentiële relatie, hebben we gekozen voor een modulaire aanpak waarbij een aparte module getraind wordt voor elk type van nominale constituent. Aangezien er nog geen vergelijkbare resultaten beschikbaar zijn voor het Nederlands hebben we onze experimenten ook uitgevoerd en geëvalueerd op de Engelse MUC-6 en MUC-7 data sets.

Toegepast op de taak van coreferentieresolutie gaan we dieper in op de methodologische aspecten die meespelen bij de toepassing van lerende systemen op natuurlijke taal. In een eerste reeks experimenten gaan we het effect na van de zogenaamde ‘bias’, de zoekheuristieken die een bepaalde leertechniek gebruikt en de manier waarop de geleerde kennis over de uit te voeren taak gerepresenteerd wordt. Daartoe evalueren we de performantie van twee lerende technieken die kunnen beschouwd worden als twee extremen in het continuüm van lerende systemen, namelijk het geheugengebaseerde systeem TIMBL en het regelinductie-

systeem RIPPER. We tonen aan dat naast de bias van het algoritme nog veel andere factoren potentieel een rol spelen in het uiteindelijke resultaat van een leerexperiment. In dit proefschrift bestuderen we het effect van de selectie van informatiebronnen, van de optimalisatie van de parameters en het effect van sampling op datasets met scheefgetrokken klassedistributies. Verder gaan we de interactie na tussen deze factoren.

In een reeks experimenten waarbij op een automatische manier relevante kennisbronnen (features) geselecteerd worden, tonen we het grote effect van featureselectie aan op de performantie van het leersysteem. We observeren ook dat de optimale featureselectie voor een bepaalde leertechniek niet kan veralgemeend worden naar andere leertechnieken. In een reeks experimenten waarbij de algoritmeparameters systematisch gevarieerd worden, tonen we verder nog aan dat de performantieverschillen binnen eenzelfde leertechniek veel groter kunnen zijn dan de performantieverschillen tussen twee of meerdere leertechnieken. Een gelijkaardige observatie kunnen we ook maken in de experimenten waarbij gekeken wordt naar de interactie tussen featureselectie en parameteroptimalisatie. Om dit soort rekenintensieve optimalisatie mogelijk te maken, wordt gebruik gemaakt van een genetisch algoritme. Deze experimenten geven ook aan dat de parameterinstellingen en de kennisbronnen die geselecteerd worden na optimalisatie niet kunnen gegeneraliseerd worden. In de experimenten waarbij de klassedistributie van de data gevarieerd wordt, tonen we aan dat beide leertechnieken zich verschillend gedragen bij scheefgetrokken klassen en dat ze ook verschillend reageren op een verandering in die distributie. Een verandering in de klasseverdeling blijkt vooral gunstig voor RIPPER. Maar ook hier kunnen we geen welbepaalde distributie aanduiden die optimaal is voor alle datasets.

Alle optimalisatie-experimenten tonen aan dat een wijziging in een van de architecturale variabelen een groot effect kan hebben op de performantie van een leer methode. Door deze conclusie komen bestaande conclusies in de literatuur op de helling te staan, omdat die vaak gebaseerd zijn op het exploreren van maar enkele punten in de experimentele ruimte. Onze studie toont aan dat er een groot risico bestaat dat andere plaatsen in de experimentele zoekruimte tot radicaal verschillende resultaten en conclusies kunnen leiden.

Op het einde van het proefschrift verlaten we het instantieniveau en concentreren we ons op de coreferentiële kettingen door de resultaten te rapporteren op de Nederlandse en Engelse testcorpora. Met het oog op een beter begrip van de fouten die begaan zijn tijdens de resoltie hebben we een kwalitatieve foutenanalyse doorgevoerd op een selectie van Engelse en Nederlandse teksten.