EXAMENSARBETE Creating a coreference solver for Swedish and German using distant supervision
STUDENT Alexander Wallin
HANDLEDARE Pierre Nugues (LTH)
EXAMINATOR Jacek Malec (LTH)

Do I refer to you?

A possible step towards solving the riddle of Swedish coreference

POPULÄRVETENSKAPLIG SAMMANFATTNING **Alexander Wallin**

It is said that coreference is difficult to explain, but easy to comprehend; everyone knows coreference, they just don't know that they do. We trained a computer to know it too!

Coreference is a relationship between two or more expressions in a text when these expressions refer to the same person or thing. Coreference solving, the identification of sets of coreferring mentions in a text, is a well-studied problem in the field of natural language processing (NLP), the computational analysis of text. As an example, consider this short text: *John drove to Judy's house. He made her dinner.* which contains four noun phrases: *John*, *Judy*, *he* and *her*. A reader would intuitively connect *John* with *he* and *Judy* with *her* and surmise that *John* cooked dinner for *Judy*. By using linguistic terms we would say that the reader has solved the text's coreference and that the links the reader previously surmised were coreferring noun phrases.

In most cases the best coreference solvers are humans, but human labour has a high resource cost and would therefore be unfeasible for most tasks; it is often better to train a computer to do the work instead, even though the results are less impressive.

To train a coreference solver, one would need to gather a large collection of text containing manually annotated coreferences. The identified coreferences are then used to train a coreference solver by comparing coreferring and non-coreferring noun phrases. Some languages are fortunate with large amounts of training data, while some languages such as Swedish have very small or nonexistent data sets for this particular task. A good rule of thumb says that the minimum training size is in the vicinity of a million words. For Swedish, there exists only one data set with 20,000 words. Besides Swedish, many languages lack large training data sets.

In the absence of a large annotated data set, distant supervision offers a possible path forward. Distant supervision in the context of our Master's thesis means that we identify identical sentences in different languages, solve coreferences for one language, and try to map them to the other language. The initial solution or the transfer may be incorrect, but given sufficiently large texts the errors would hopefully be negligible.

The goal for our Master thesis is the creation of coreference solvers for the Swedish and German languages using this method.

Although the methods we describe have been used with some success in other languages, to the best of our knowledge, we are the first to create a coreference solver for Swedish using this technique.

We hope our results will pave the way for the creation of coreference solvers competitive with the current state of the art achieved by supervised training techniques.