

**EXAMENSARBETE** HERD - A Named Entity Recognizer and Disambiguator

**STUDENT** Anton Södergren

**HANDLEDARE** Pierre Nugues (LTH)

**EXAMINATOR** Elin Anna Topp (LTH)

# Namnigenkänning och länkning i naturligt språk

---

POPULÄRVETENSKAPLIG SAMMANFATTNING **Anton Södergren**

---

För att samla stora mängder information om personer, platser och organisationer, behöver vi kunna analysera vanliga texter skrivna i naturligt språk. Detta arbete bidrar till det, genom att känna igen namn och länka dem till rätt Wikipedia-artikel.

För att kunna skapa system som kan svara på frågor, eller ge rätt svar vid sökningar, behöver vi analysera text: den största källan av information som finns tillgänglig. Eftersom det är en ostrukturerad och tvetydig informationskälla, behöver vi först transformera den till något som är lättare för datorer att förstå. Det här arbetet ämnar att göra text mer strukturerad, genom att känna igen namn i text, och länka varje namn till en artikel på Wikipedia, om en sådan artikel finns.

Det finns två stora svårigheter med namnigenkänning. Den första är att varje namn har flera olika möjliga lösningar. Det finns till exempel 21 personer vid namn "Michael Jackson" kända nog att ha sin egen Wikipedia-sida, och tusentals fler som inte har det. Den andra svårigheten är att avgöra om ett stycke text innehåller ett namn, vilket är lättare att förstå givet att det finns en artist, en amerikansk by, en telefonoperatör och en kinesisk bank som alla heter "Tomato", vilket oftast inte är ett namn alls.

Mitt examensarbete har fokuserat på att skapa ett system för att känna igen namn och länka dem till en unik identifierare. Jag har skapat en algoritm som samlar idéer från olika state-of-the-art system, och utvärderat dem. Systemet skulle vara snabbt nog att kunna köras på hela Wikipedia inom en rimlig tids-

ram, och i kontrast med många andra system skulle det vara flerspråkigt. Dess effektivitet mättes på samma vis som ERD'14, en stor tävling där forskargrupper tävlade med sina system på samma dataset.

Systemet jag har byggt baseras på att analysera alla länkar mellan olika sidor i Wikipedia, och sedan räkna hur ofta en viss fras leder till en viss sida. På så vis kan vi samla alla olika namn och smeknamn en viss entitet har, och dessutom avgöra vilken länk som är vanligast för varje fras. Denna data laddas in i en databas och vid läsning av nya texter markeras alla fraser som finns i databasen.

Vilka fraser som ska behållas avgörs i flera steg. Det första är att med några enkla regler ta bort det minst troliga. Sedan körs en iterativ variant av algoritmen PageRank. Den försöker avgöra hur mycket två olika namn har gemensamt, genom att titta på om de har länkat till varandra i Wikipedia, eller nämnts i samma artikel tidigare. Systemet väljer sedan vem som menas med en fras, genom att ta den artikel med flest antal länkar för den frasen i Wikipedia.

Systemet nådde ett F1-resultat, vilket är ett genomsnitt av kvalitén för metoden, på 70% för engelska, vilket hade placerat det på en sjätteplats i ERD'14. Systemet finns även tillgängligt för svenska och franska.