Learning natural language morphology from a raw text

Jackson Lee (Graduate student in linguistics)
John Goldsmith (Faculty advisor)
Simon Jacobs (Consultant at RCC)
{jsllee, goldsmith, sdjacobs}@uchicago.edu

THE UNIVERSITY OF CHICAGO

Goal

Develop an unsupervised, language-independent system that takes:

Input – raw text

> The Fulton County Grand Jury said Friday an investigation of Atlanta's...

...and generates:

Output – morphological paradigms

talk	talked	talking	talks	etc.
move	moved	moving	moves	

Resources

Data

– Brown corpus (1 million words)
– Google n-gram corpus (from Google Books, 4 billion words)

RCC

– storage, memory, cluster computing
– data visualization

Approach

From word trigrams to word contexts
⇒ **word context visualization**

Computing word similarity
⇒ **word manifolds**

Inducing paradigm tables
⇒ **word manifolds with paradigms**

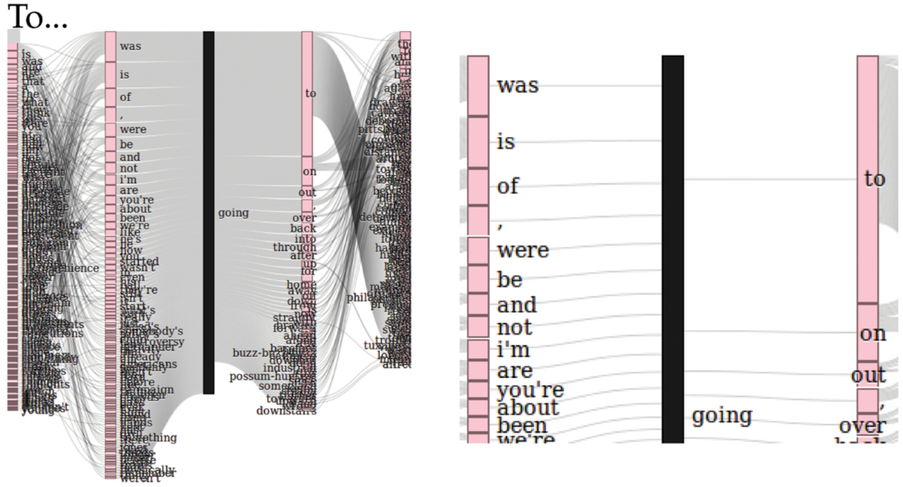
Word context visualization

From **trigrams**:

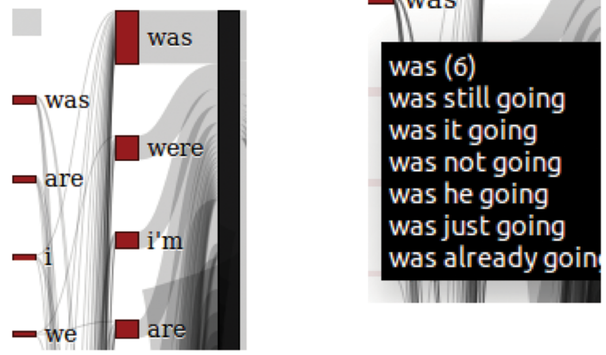
(e.g., *going*)

was going after
going to allow
was not going

etc.



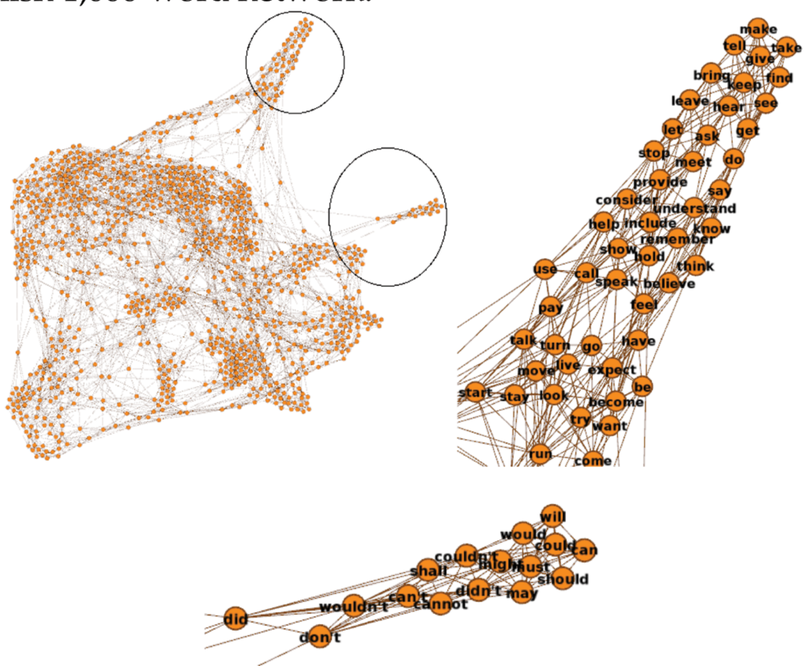
Context words
at multiple positions:



Word manifolds

A graph-theoretic approach to computing distributional similarities among words (Goldsmith & Wang 2012)

English 1,000-word network:



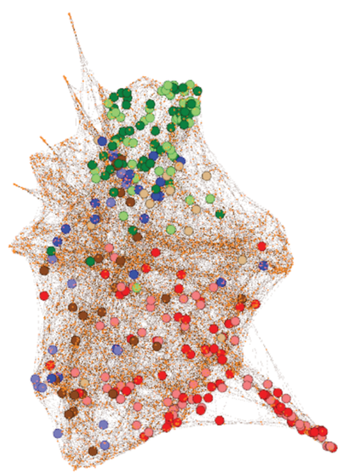
Word manifolds with paradigms

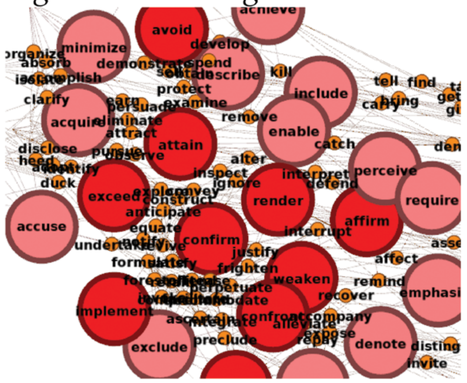
Linguistica (Goldsmith 2001): Inducing stems + affixes

stems	affix pattern
jump, walk, ...	Ø-ed-ing-s
lov, mov, ...	e-ed-es-ing

But no alignment across affix patterns

Solution: Combine *Linguistica* and word manifolds
The match between *Ø-ed-ing-s* and *e-ed-ing-es*





affix pattern 1: Ø ed ing s (darker)
affix pattern 2: e ed ing es (lighter)