COGNITIVE COMPUTATION GROUP
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Cognitive Computation Group

Natural Language Processing Tools Overview
October 28, 2014

http://cogcomp.cs.illinois.edu

MULTIMODAL INFORMATION ACCESS & SYNTHESIS mias

Outline

- CCG NLP Tools for enriching text
- Illinois NLP Curator: managing Annotators
- IllinoisCloudNLP: text analytics in the cloud
- Comparators: computing text similarity
- Learning Based Java: integrating machine learning directly into applications

CCG NLP TOOLS

Available from CCG

- Tokenization/Sentence Splitting
- Part Of Speech
- Chunking
- Named Entity Recognition
- Coreference
- Semantic Role Labeling
- Wikifier
- Hierarchical Dataless Classifier

Tokenization and Sentence Segmentation

- Given a document, find the sentence and token boundaries

> The police chased Mr. Smith of Pink Forest, Fla. all the way to Bethesda, where he lived. Smith had escaped after a shoot-out at his workplace, Machinery Inc.

- Why?
  - Word counts may be important features
  - Words may themselves be the object you want to classify
  - "lived." and "lived" should give the same information
  - different analyses need to align if you want to leverage multiple annotators from different sources/tasks

Tokenization and Sentence Segmentation ctd.

- Believe it or not, this is an open problem
- No single standard for token-level segmentation
  - e.g. "American-led" vs. "American - led"?
  - e.g. "$ 32 M" vs "$32 M" and "$32M"?
- Different tasks may use different standards
- No wildly successful sentence segmenter exists (see the excerpts in news aggregators for some nice errors)
- Noisier text (e.g. online consumer reviews) => poorer performance (for reasons like inconsistent capitalization)
- LBJava distribution includes the Illinois tokenizer and sentence segmenter

Part of Speech (POS)

- Allows simple abstraction for pattern detection

POS	DT	NN	VBD	PP	DT	JJ	NN
Word	The	boy	stood	on	the	burning	deck

POS	DT	NN	VBD	PP	DT	JJ	NN
Word	A	boy	rode	on	a	red	bicycle

- Disambiguate a target, e.g.

  "make (a cake)" vs. "make (of car)"

- Specify more abstract patterns,

  e.g. Noun Phrase: (DT JJ* NN)

- Specify context in abstract way
  - e.g. "DT boy VBX" for "actions boys do"
  - This expression will catch "a boy cried", "some boy ran", …

Chunking

- Identifies phrase-level constituents in sentences

  [NP Boris] [ADVP regretfully] [VP told] [NP his wife]
  [SBAR that] [NP their child] [VP could not attend]
  [NP night school] [PP without] [NP permission] .

- Useful for filtering: identify e.g. only noun phrases, or only verb phrases
  - Groups modifiers with heads
- Used as source of features, e.g. distance (abstracts away determiners, adjectives, for example), sequence,…
  - More efficient to compute than full syntactic parse
  - Applications in Information Extraction, e.g. Term Extraction

Named Entity Recognition

- Identifies and classifies strings of characters representing proper nouns:

In **[LOC South Ossetia]** , **[ORG Human Rights Watch]** confirmed that a cluster strike in the center of the city of **[LOC Gori]** killed at least eight civilians, including **[MISC Dutch]** journalist **[PER Stan Storimans]**. **[MISC Israeli]** journalist **[PER Zadok Yehezkeli]** was among the injured.

Coreference

- Identify all phrases that refer to each entity of interest – i.e., group mentions of concepts

  After checking in with pilot **[Buzz Aldrin]**, **[Neil]** radioed to **[earth]**. With a serious look on **[his]** face, **[the 38-year-old civilian commander]** said the famous words, "**[the Eagle]** has landed"."

- The Named Entity recognizer only gets us part-way…
- …if we ask, "what actions did Neil Armstrong perform?", we will miss many instances (e.g. "He said…")
- Coreference resolver **abstracts over different ways of referring to the same person**
  - Useful in feature extraction, information extraction

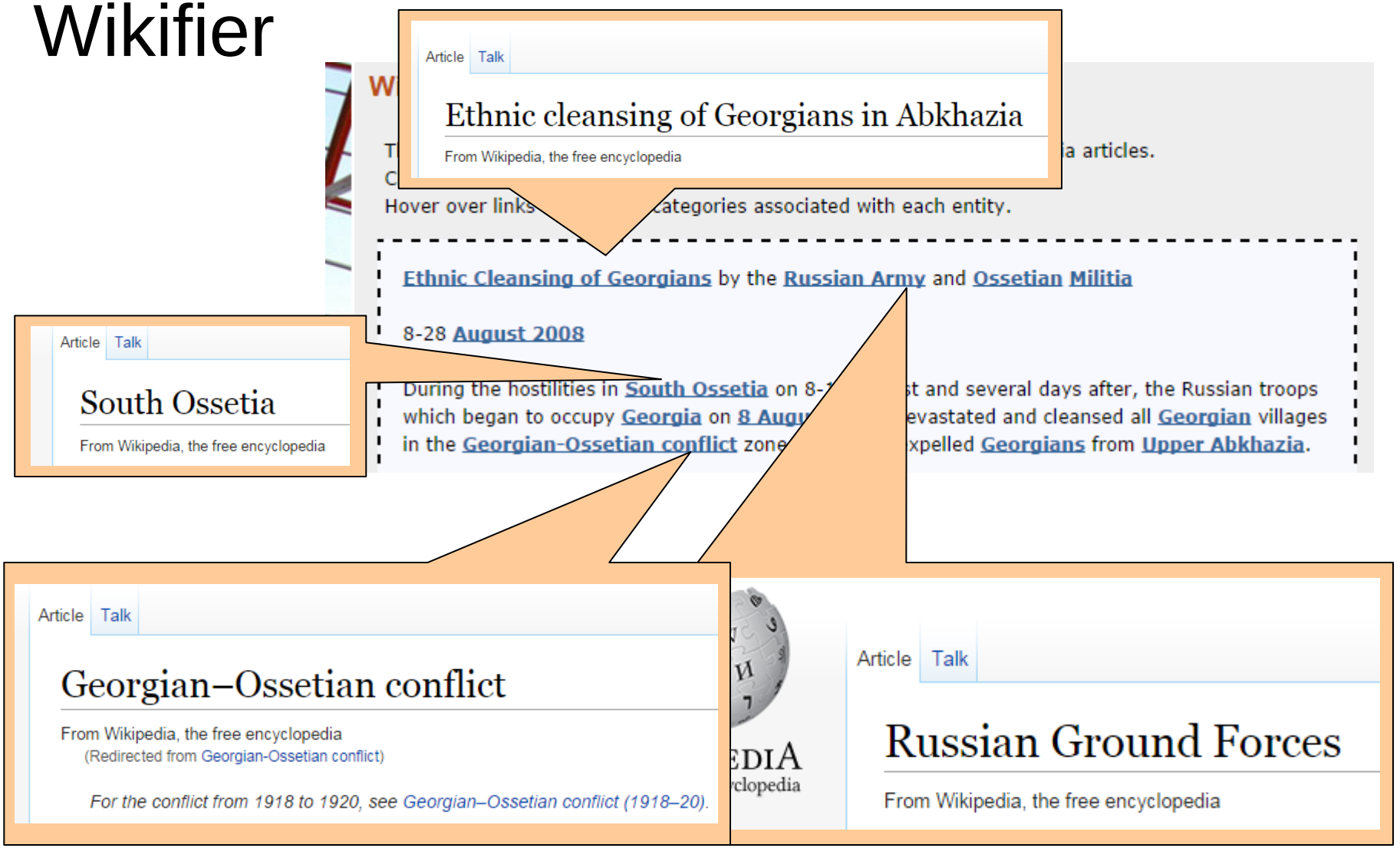
Semantic Role Labeler

Output:

	NE	SRL	Nom	trip	Preposition	Preposition	Preposition
Bart	Person	rider [A0]					
rode		V: ride.01			Governor		Governor
his		steed [A1]					
bike						Governor	
from		direction [AM-DIR]				Source (from)	
Springfield	Geo-political Entity					Object	
to		direction [AM-DIR]			Destination (to)		
Shelbyville	Geo-political Entity				Object		
on		temporal [AM-TMP]					Temporal (on)
Tuesday	Date						Object
.							
Bart	Person			traveller [A0]			
's							
trip			trip 01				
to				destination or path [A1]			
Shelbyville	Geo-political Entity						
was							
on							
Tuesday	Date						
.							

- SRL reveals relations and arguments in the sentence (where relations are expressed as verbs)
- Cannot abstract over variability of expressing the relations – e.g. kill vs. murder vs. slay…

Wikifier


Article
Talk
Ethnic cleansing of Georgians in Abkhazia
From Wikipedia, the free encyclopedia
Hover over links
categories associated with each entity.
Ethnic Cleansing of Georgians by the Russian Army and Ossetian Militia
8-28 August 2008
During the hostilities in South Ossetia on 8-
st and several days after, the Russian troops
which began to occupy Georgia on 8 Augu
evastated and cleansed all Georgian villages
in the Georgian-Ossetian conflict zone
xpelled Georgians from Upper Abkhazia.
Article
Talk
South Ossetia
From Wikipedia, the free encyclopedia
Article
Talk
Georgian–Ossetian conflict
From Wikipedia, the free encyclopedia
(Redirected from Georgian-Ossetian conflict)
For the conflict from 1918 to 1920, see Georgian–Ossetian conflict (1918–20).
Article
Talk
Russian Ground Forces
From Wikipedia, the free encyclopedia

Performance

Tool	Publictn	Dataset	CCG	Best Other
Co-reference	EMNLP '13	OntoNotes 5	63.30 (avg. of MUC, B^3, CEAF)	63.37*
Named Entity	ACL '11	CoNLL '03	90.36	90.90*
Wikifier	EMNLP '13	Custom	87.12 (avg. over 4 data sets)	76.30
SRL (Verb)	CoNLL '05	WSJ+Brown	**77.92**	77.30
SRL (Prep)	EMNLP '11	WSJ 23	67.82	-
Dataless	AAAI '14	20NG (unsup)	68.2/83.7**	59.5

*could not find online release of software
** second result uses bootstrapping