Representing Text Chunks

Erik F. Tjong Kim Sang
Center for Dutch Language and Speech
Universitaire Instelling Antwerpen
Belgium
erikt@uia.ua.ac.be

Jorn Veenstra
Computational Linguistics
Tilburg University
The Netherlands
veenstra@kub.nl

overhead sheets: http://lcg-www.uia.ac.be/˜erikt/talks/

TMR Network

Learning Computational Grammars

Goal

Apply machine learning techniques for recognizing the structure of noun phrases (NPs).

Steps

1. Recognize NP boundaries.
2. Discover syntactic structure within NPs.
3. Find out semantic roles of NP constituents.

NP Chunking

Recognize non-recursive noun phrases (baseNPs): NPs that do not contain another NP.

Ramshaw & Marcus (WVLC95) have used Transformation-Based Learning for training an NP chunker on a subset of the Wall Street Journal corpus.

Their NP chunker achieved a precision and recall rates of approximately 92% while recognizing baseNPs from section 20 of this corpus.

Data representation

RM95: Text chunking can be represented as a tagging task by using three tags: I, O and B

- In $[_N$ early trading $_N]$ in $[_N$ Hong Kong $_N]$ $[_N$ Monday $_N]$, $[_N$ gold $_N]$ was quoted at $[_N$ \$ 366.50 $_N]$ $[_N$ an ounce $_N]$.
- In/O early/I trading/I in/O Hong/I Kong/I Monday/B ,/O gold/I was/O quoted/O at/O $/I 366.50/I an/B ounce/I ./O

Advantage: tags are less dependent on each other than brackets. Problems can be solved locally.

Alternative data representation formats

IOB1	O	I	I	O	I	I	B	O	I	O	O	O	I	I	B	I	O
IOB2	O	B	I	O	B	I	B	O	B	O	O	O	B	I	B	I	O
IOE1	O	I	I	O	I	E	I	O	I	O	O	O	I	E	I	I	O
IOE2	O	I	E	O	I	E	E	O	E	O	O	O	I	E	I	E	O
IO	O	I	I	O	I	I	I	O	I	O	O	O	I	I	I	I	O
[	.	[	.	.	[	.	[	.	[	.	.	.	[	.	[	.	.
]	.	.	]	.	.	]	]	.	]	.	.	.	.	]	.	]	.

Experiment setup

Memory-based learning (IB1-IG) has been used for performing five-fold cross-validation experiments on section 15 of the Wall Street Journal corpus as prepared by RM95.

Each experiment included four parts. Each part examined different parameter settings.

Results have been measured by examining an average of recall and precision rates:
F = (2×precision×recall)/(precision+recall)

The best parameter settings have been used for processing the complete RM95 data set (sections 15-18 from WSJ as training material and section 20 for testing).

Part 1: determining context size

Example: In/IN early/JJ trading/NN $\rightarrow$ I

IOB1	R=0	R=1	R=2	R=3	R=4
L=0	81.31	85.15	84.81	84.19	83.87
L=1	87.09	89.12	88.74	88.42	88.22
L=2	86.33	89.17	88.87	88.76	88.59
L=3	85.83	88.96	88.64	88.60	88.43
L=4	85.78	88.65	88.43	88.41	88.34

	Context	F
IOB1	L=2/R=1	89.17
IOB2	L=2/R=1	88.76
IOE1	L=1/R=2	88.67
IOE2	L=2/R=2	89.01
[+]	L=2/R=1 + L=0/R=2	89.32
[+ IO	L=2/R=0 + L=1/R=1	89.43
IO +]	L=1/R=1 + L=0/R=2	89.42

Part 2: determining IOB context size

In/IN/O early/JJ trading/NN/I $\rightarrow$ I

IOB1	R=0	R=1	R=2	R=3
L=0	89.17	89.77	89.65	89.57
L=1	89.68	90.11	90.12	90.02
L=2	89.64	90.11	89.99	89.98
L=3	89.51	89.96	89.95	89.92

	Context	IOB Context	F
IOB1	L=2/R=1	1/2	90.12
IOB2	L=2/R=1	1/0	89.30
IOE1	L=1/R=2	1/2	89.55
IOE2	L=1/R=2	0/1	89.73
[+]	L=2/R=1 + L=0/R=2	0/0 + 0/0	89.32
[+ IO	L=2/R=0 + L=1/R=1	0/0 + 1/1	89.78
IO +]	L=1/R=1 + L=0/R=2	1/1 + 0/0	89.86

Part 3: combine some part 1 results

RM95 use transformation rules with different context sizes. It might help if we combine different results from part 1 in the second processing step.

IOB1	R=0	R=1	R=2	R=3	R=4
L=0	81.31	85.15	84.81	84.19	83.87
L=1	87.09	89.12	88.74	88.42	88.22
L=2	86.33	89.17	88.87	88.76	88.59
L=3	85.83	88.96	88.64	88.60	88.43
L=4	85.78	88.65	88.43	88.41	88.34

	Context	part 1	F
IOB1	2/1	0/0 1/1 2/2 3/3	90.53
IOB2	2/1	2/1	89.30
IOE1	1/2	0/0 1/1 2/2 3/3	90.03
IOE2	1/2	1/2	89.73
[+]	2/1 + 0/2	- + -	89.32
[+ IO	2/0 + 1/1	- + 0/1 1/2 2/3 3/4	89.91
IO +]	1/1 + 0/2	0/1 1/2 2/3 3/4 + -	90.03

Part 4: change k

In IB1-IG, k is the number of nearest neighbors that are considered when determining the most probable output.

Daelemans, Van den Bosch and Zavrel (ML99): abstraction (k>1) is harmful in language learning.

But not for NP chunking?

	word/POS	$F_{\beta=1}$
IOB1	3/3(k=3)	90.89 ± 0.63
IOB2	3/3(k=3)	89.72 ± 0.79
IOE1	2/3(k=3)	90.12 ± 0.27
IOE2	2/3(k=3)	90.02 ± 0.48
[+]	4/3(3) + 4/4(3)	90.08 ± 0.57
[+ IO	4/3(3) + 3/3(3)	90.35 ± 0.75
IO +]	3/3(3) + 2/3(3)	90.23 ± 0.73

Processing the RM95 data set

	accuracy	precision	recall	$F_{\beta=1}$
IOB1	97.58%	92.50%	92.25%	92.37
IOB2	96.50%	91.24%	92.32%	91.78
IOE1	97.58%	92.41%	92.04%	92.23
IOE2	96.77%	91.93%	92.46%	92.20
[+]	-	93.66%	90.81%	92.22
[+ IO	-	91.47%	92.61%	92.04
IO +]	-	91.25%	92.54%	91.89

Results of earlier work

	accuracy	precision	recall	F
RM95	97.37%	91.80%	92.27%	92.03
Vee98	97.2%	89.0%	94.3%	91.6
ADK98	-	91.6%	91.6%	91.6
CP98	-	90.7%	91.1%	90.9

Post-paper work

We have discovered a bug in the software that combined bracket structures to complete NPs. It generated NPs at places where no closing bracket was present.

Rates after software correction

15	accuracy	precision	recall	$F_{\beta=1}$
[+]	-	93.21%	87.94%	90.50
IOB1	97.16%	91.08%	90.69%	90.89

20	accuracy	precision	recall	$F_{\beta=1}$
[+]	-	94.72%	90.78%	92.71
IOB1	97.58%	92.50%	92.25%	92.37
RM95	97.37%	91.80%	92.27%	92.03

00	accuracy	precision	recall	$F_{\beta=1}$
[+]	-	95.86%	92.55%	94.18
IOB1	98.04%	93.71%	93.90%	93.81
RM95	97.8%	93.1%	93.5%	93.3

Examples of chunking errors

First ten errors in WSJ section 20, processed with [+].

- , torque box , fixed leading edges
- [an aerospace] , [electronics] , automotive and graphics concern
- [SHEARSON LEHMAN HUTTON Inc] .
- [his previous real-estate investment] and asset-management duties
- [development and property management]
- the_DT savings_NNS bank_VBP
- [its total loan] and [real estate reserves]
- [the_DT savings_NNS] bank_VBP
- [takeover] stock speculators
- [shares] of UAL , [United]

Concluding remarks

- For baseNP recognition, representing text chunking as a bracketing task gives as least as good results as representing text chunking as a tagging task.
- With memory-based learning good results with a bracketing representation are easier to obtain than with a tagging representation.
- From the four tagging schemes that have been examined (IOB1, IOB2, IOE1 and IOE2) the first seems to be the best although the performance differences with the other three are small.

Future work

- Apply this paradigm to recognizing arbitrary noun phrases.

Address

address:	Center for Dutch Language and Speech Universitaire Instelling Antwerpen Universiteitsplein 1 B-2610 Wilrijk Belgium
phone:	+31 3 8202765
e-mail:	erikt@uia.ua.ac.be
www:	http://lcg-www.uia.ac.be/˜erikt/

Overhead sheets

http://lcg-www.uia.ac.be/˜erikt/talks/

Software (TiMBL)

http://ilk.kub.nl/

Data

ftp://ftp.cis.upenn.edu/pub/chunker/