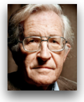
Discriminative · Lexical Semantic Segmentation · with · Gaps: Running the MWE Gamut

Multiword expressions (MWEs) are **diverse** and **collectively frequent** in English. We train a supervised discriminative **sequence model** on a new **annotated corpus** to identify heterogenous MWEs in context, giving a **lexical semantic segmentation** of the sentence. We extend shallow chunking to capture **gappy** (discontinuous) expressions.

Multiword Expressions

Definition: ≥2 space-separated words whose combination is idiosyncratic in *form*, *function*, and/or *distribution*.
Diverse syntax and semantics:

Noam Chomsky



daddy longlegs, hot dog
dry out
depend on, come across
pay attention (to)
put up with, give in (to)
under the weather
cut and dry
in spite of
pick up where __ left off
easy as pie
You're welcome.
To each his own.
The structure of this paper is as follows.

They gave_ me _the_run_around and missing paperwork only to call_back to tell me someone else wanted her and I would need to come_in and put_down~ a ~deposit .

Labeled Data

CMWE, a text corpus comprehensively annotated with **multiword expressions**
(Schneider et al., LREC 2014)

- ✦ 3,500 manually annotated MWE instances in 3,800 sentences (55k words) of English web reviews
- ✴ fully heterogeneous MWEs
- ✴ shallow groupings, allowing gaps
- ✴ strong (put_down) vs. weak (put_down~deposit)

Gappy Sequence Tagging

Problem: Identify MWEs as chunks with possible **gaps,** so as to apply **tagging**.
Solution: Double the BIO tagset to encode gap status in the state space. Full model: 8 tags

		token in gap	
	O	o	
token part of MWE	B	b	
	Ī	ī	strong continuation
	Ĩ	ĩ	weak continuation

O	O	B	Ī	O	B	Ī	o	Ĩ
need	to	come_	in	and	put_	down~	a	~deposit

Link-Based Evaluation

Gives partial credit for **partial overlap** between predicted and gold MWEs. See paper for details.

Experiments

Preprocessing: POS tag (retrained TweetNLP tagger on rest of English Web Treebank)
Model: First-order **structured perceptron** tagger (Collins, 2002) with **recall-oriented cost** to balance recall and precision (Mohit et al., 2012)
Features:

- ✴ Basic features (summarized below)
- ✴ MWE lexicon match
  - MWE lexicons extracted from WordNet, SemCor, Prague Czech-English Treebank, SAID, WikiMwe, Wiktionary, and other lists
- ✴ Brown clusters from Yelp Academic Dataset

Baseline: Match lemmas against lexicons, predict the segmentation with fewest total expressions.

Basic features adapted from Constant et al. (2012):

- **word:** current & context, unigrams & bigrams
- **POS:** current & context, unigrams & bigrams
- capitalization; word shape
- prefixes, suffixes up to 4 characters
- has digit; non-alphanumeric characters
- lemma + context lemma if one is a V and the other is ∊ {N, V, Adj., Adv., Prep., Part.}

Results

supervised model ≫ non-statistical baseline; lexicon matching features help (of {0,2,6,10} lexicons to consult, 6 is best); and:

configuration	iters	cost	params	P	R	F_1
base model	5	—	1,765k	69.27	50.49	58.35
+ recall cost	4	150	1,765k	61.09	57.94	59.41
+ clusters	3	100	2,146k	63.98	55.51	59.39
+ oracle POS	4	100	2,145k	66.19	59.35	62.53

Nathan Schneider · Emily Danchik · Chris Dyer · Noah A. Smith @ Carnegie Mellon University

TACL 2014
PARSEME Malta 2015

corpus + **tagger** @ http://www.ark.cs.cmu.edu/LexSem/

Carnegie Mellon