WHUNlp at SemEval-2016 Task DiMSUM: A Pilot Study in Detecting Minimal Semantic Units and their Meanings using Supervised Models

Xin Tang and **Fei Li** and **Donghong Ji**
Computer School, Wuhan University, Wuhan, China
1648706227@qq.com, lifei_csnlp and dhji@whu.edu.cn

Abstract

This paper describes our approach towards the SemEval-2016 Task 10: Detecting Minimal Semantic Units and their Meanings (DiMSUM). We consider that the two problems are similar to multiword expression detection and supersense tagging, respectively. The former problem is formalized as a sequence labeling problem solved by first-order CRFs, and the latter one is formalized as a classification problem solved by Maximum Entropy Algorithm. To carry out our pilot study quickly, we extract some simple features such as words or part-of-speech tags from the training set, and avoid using external resources such as WordNet or Brown clusters which are allowed in the supervised closed condition. Experimental results show that much further work on feature engineering and model optimization needs to be explored.

1 Introduction

In the community of natural language processing, multiword expressions (MWEs) detection (Schneider et al., 2014b; Schneider et al., 2014a) and supersense tagging (Ciaramita and Johnson, 2003; Ciaramita and Altun, 2006) have received much research attention due to their various applications such as syntactic parsing (Candito and Constant, 2014; Bengoetxea et al., 2014), semantic parsing (Banarescu et al., 2013), and machine translation (Carpuat and Diab, 2010). However, not much attention has been paid to the relationship between MWEs and supersenses (Piao et al., 2005; Schneider and Smith, 2015).

Input: Security$_{NOUN}$ increased$_{VERB}$ in$_{ADP}$
Mumbai$_{PROPN}$ amid$_{ADP}$
terror$_{NOUN}$ threats$_{NOUN}$
ahead$_{ADP}$ of$_{ADP}$ Ganeshotsav$_{PROPN}$

Output: Security$_{n.state}$ increased$_{v.change}$ in
Mumbai$_{n.location}$ amid
terror_threats$_{n.communication}$
ahead of Ganeshotsav$_{n.event}$

Figure 1: An DiMSUM Example. Given a tokenized and POS-tagged sentence, outputs will be a representation annotated with MWEs and supersenses. Noun and verb supersenses start with "n." and "v.", respectively. "_" joins tokens within a MWE.

The DiMSUM shared task (Schneider et al., 2016) at SemEval 2016 aims to predict a broad-coverage representation of lexical semantics giving an English sentence. This representation consists of two facets: a segmentation into minimal semantic units, and a labeling of some of those units with semantic classes. Based on the task descriptions, we consider the concepts of minimal semantic units and semantic classes are identical to those of MWEs and supersenses, respectively. Figure 1 shows an input example and its corresponding outputs of DiMSUM task.

Prior work on MWE detection using unsupervised methods includes lexicon lookup (Bejček et al., 2013), statistical association measures (Ramisch et al., 2012), parallel corpora (Tsvetkov and Wintner, 2010), or hybrid methods (Tsvetkov and Wintner, 2011). More sophisticated methods use supervised techniques such as conditional random fields (CRFs) (Shigeto et al., 2013; Constant et al., 2012; Vincze et al., 2013) or structured perceptron

Proceedings of SemEval-2016, pages 918–924,
San Diego, California, June 16-17, 2016. ©2016 Association for Computational Linguistics

(Schneider et al., 2014a), and usually achieve better performance. Compared to most aforementioned systems, MWEs in the DiMSUM task may be not contiguous or restricted by syntactic construction, which increases the detection difficulty.

Supersense tagging has been studied on diverse languages such as English (Ciaramita and Johnson, 2003; Ciaramita and Altun, 2006; Johannsen et al., 2014; Schneider and Smith, 2015), Italian (Attardi et al., 2010), Chinese (Qiu et al., 2011) and Arabic (Schneider et al., 2012). It is usually formalized as a multi-classification problem solved by supervised approaches such as perceptron. In the DiMSUM task, both single-word and multiword expressions that holistically function as noun or verb, can be considered as units for supersense tagging.

Following prior work using supervised approaches, we divide DiMSUM task into two subtasks: first, MWEs detection is treated as a sequence labeling task using first-order CRFs; second, supersense tagging is treated as a multi-classification task using Maximum Entropy Algorithm. We focus on the supervised closed condition, so only the training set are used for training both submodels separately. Then results generated on the test set are submitted for official evaluation. The evaluation results show that our system performance are not as good as those of other teams, since we leverage only some simple features such as words, POS, etc. Syntactic features and semantic resources such as WordNet (Miller, 1995) and Brown clusters (Brown et al., 1992) are not used. This suggests that further work needs to be done on feature engineering and model optimization.

2 Multiword Expression Detection

In the training set, sentences are tokenized into words, and every word has been annotated with POS and lemma. We formalize MWE detection as a sequence labeling problem, so Mallet (McCallum, 2002), Off-the-shelf implementation of CRFs, is used to handle this task. All the labels of our CRF model, extracted from the training set, are listed as follows:

- O, which indicates that the current word does not belong to a MWE.
- B, which indicates that the current word is the first token of a MWE.
- I, which denotes that the current word continues a MWE.
- o, which indicates that the current word does not belong to a MWE, but inside the gap of another MWE.
- b, which denotes that the current word is the first token of a MWE and inside the gap of another MWE.
- i, which indicates that the current word continues a MWE and inside the gap of another MWE.

1	current word w_i
2	whether current word w_i is in the beginning of a sentence
3	whether current word w_i is in the end of a sentence
4	previous word w_{i-1}
5	next word w_{i+1}
6	POS of current word t_i
7	POS of previous word t_{i-1}
8	POS of next word t_{i+1}
9	whether current word w_i is the only word of a sentence

Table 1: Feature templates for MWE Detection.

Compared with prior work on MWE detection, one difference in DiMSUM is that gaps may exist in MWEs, which increases difficulty to recognize them correctly. For example, given “Bramen Honda was a bit of a hassle .” as input, the output labels will be “Bramen$_B$ Honda$_I$ was$_B$ a$_b$ bit$_i$ of$_o$ a$_I$ hassle$_I$.$_O$”. Tag “B” and tag “I” can be discontinuous and there may be several “b”, “i” or “o” between them.

In order to implement a system for our pilot study quickly, we only use some simple features which are shown in Table 1. The values of these features are discrete, namely 0 or 1. The motivation of Feature 7 and 8 is to help our model to generate correct label sequences which satisfy some constraints. For example, only “B” and “O” can be located in the beginning of a sentence, and “b”, “i” and “o” cannot be in the end of a sentence. Feature 9 is added since most of words which is the only one of a sentence, are tagged as “O” in the training set.

3 Supersense Tagging

We treat supersense tagging as a multi-classification problem using Maximum Entropy Algorithm, and

noun	verb
act, animal, artifact attribute, body, cognition communication, event food, group, location motive, natural_object other, person, phenomenon plant, possession, process quantity, relation, shape state, substance, time feeling	body, change cognition, social communication competition consumption contact, weather creation, motion emotion, stative perception possession

Table 2: Supersense Categories.

1	$w_1w_2...w_n$ of *mwe*$_i$ or *swe*$_i$
2	$w_1w_2...w_n$ of *mwe*$_{i-1}$ or *swe*$_{i-1}$
3	$w_1w_2...w_n$ of *mwe*$_{i+1}$ or *swe*$_{i+1}$
4	$t_1t_2...t_n$ of *mwe*$_i$ or *swe*$_i$
5	$t_1t_2...t_n$ of *mwe*$_{i-1}$ or *swe*$_{i-1}$
6	$t_1t_2...t_n$ of *mwe*$_{i+1}$ or *swe*$_{i+1}$

Table 3: Feature templates for Supersense Tagging. *swe* or *mwe* denotes a single or multiple word expression. *mwe*$_i$, *mwe*$_{i-1}$ and *mwe*$_{i+1}$ denote current, previous and next multiple word expressions, respectively. $w_1w_2...w_n$ and $t_1t_2...t_n$ denote word and POS combinations, respectively.

Mallet is also used to implement our supersense tagging subsystem. Based on the task description, single-word or multiword expressions can receive supersenses, so both of them are treated as classification units. This suggests that supersense tagging does not totally depend on the results of MWE detection. According to Schneider and Smith (2015), supersense categories are listed as Table 2.

Given "Bramen$_B$ Honda$_I$ was$_B$ a$_b$ bit$_i$ of$_o$ a$_I$ hassle$_I$.$_O$" as input, our model will firstly transform it into classification units based on the labels, "[Bramen Honda], [was a hassle], [a bit], [of], [.]". Since the span of "[was a hassle]" includes "[a bit]" and "[of]", it is located before "[a bit]" and "[of]". Then these units will be classified into supersense categories and an empty class which receives the units which do not belong to any category.

Supersense tagging features are shown in Table 3. The values of these features are also discrete.

Features	**F$_1$**
Baseline	28.9
+4	36.5
+5	31.4
+6	51.5
+7	50.4
+8	19.9
+9	34.8

(a)

Features	**F$_1$**
Baseline	50.3
+2	42.2
+3	40.1
+4	55.5
+5	50.7
+6	52.2

(b)

Table 4: Subtable (a) and (b) dennote contributions of features in Table 1 and 3, respectively. "+" denotes that only the feature in the current line is added. The numbers in the first column correspond to the ones in Table 1 and 3, respectively.

4 Experiments

4.1 Experimental Settings

There are three conditions in the DiMSUM[1] shared task at SemEval 2016. In the supervised closed condition, only the training set, WordNet and Brown clusters are allowed to be used. In the semi-supervised closed condition, all of the above are permitted, plus the Yelp Academic Dataset. In the open condition, all available resources can be used. We carry out our experiments according to the demands in the supervised closed condition, but WordNet and Brown clusters are not used. The test set consists of 16,500 words in 1,000 English sentences which are drawn from the following sources: reviews from the TrustPilot corpus (Hovy et al., 2015), tweets from the Tweebank corpus (Kong et al., 2014), TED talks from the WIT3 archive (Cettolo et al., 2012).

During the development period, we split 30% data from the training set as our development set and use the remainder for training. The maximum training iteration is set as 500. The evaluation scripts (v1.5) released by task organizers are used for tuning parameters and features. During the official evaluation period, we use all the training set to train our CRF and Maximum Entropy models, and prediction results on the test set are submitted.

4.2 Development Results

The contributions of features are shown in Table 4. The baseline in Table 4a does not only use Feature 1

[1] http://dimsum16.github.io/

Team	Condition	μ-M	μ-S	μ-C	Tw-M	Tw-S	Tw-C	R-M	R-S	R-C	TED-M	TED-S	TED-C	Macro-C
ICL-HD	open	56.66	57.55	57.41	59.49	55.99	56.63	53.37	57.66	56.98	57.14	60.06	59.71	57.77
VectorWeavers	open	38.49	51.62	49.77	39.32	51.70	49.74	36.18	51.36	49.25	42.76	52	50.82	49.94
UW-CSE	open	57.24	57.64	57.57	61.09	57.46	58.18	54.80	57	56.61	53.48	59.17	58.33	57.71
BCED	open	13.48	51.93	46.64	15.50	51.11	45.44	8.68	51.98	46.15	20.11	53.28	49.81	47.13
BCED	semi-closed	13.46	51.11	45.86	15.76	49.95	44.42	9.07	52	46.19	18.28	51.40	47.90	46.17
UFRGS	super-closed	51.48	49.98	50.22	51.16	49.20	49.54	49.57	50.93	50.71	56.76	49.61	50.57	50.27
WHUNlp	super-closed	**30.98**	**25.14**	**25.76**	**34.18**	**24.63**	**25.87**	**26.39**	**25.82**	**25.86**	**33.44**	**24.68**	**25.39**	**25.71**
UW-CSE	super-closed	53.93	57.47	56.88	54.48	56.82	56.38	53.96	57.19	56.66	52.35	59.11	58.26	57.10
BCED	super-closed	8.20	51.29	45.47	6.34	49.66	42.99	7.05	52.68	46.57	16.30	51.44	47.82	45.79
UW-CSE	open(late)	56.71	57.72	57.54	61.96	57.65	58.51	52.09	57.22	56.31	54.09	58.78	58.16	57.66

Table 5: Official Evaluation Results (in %) of DiMSUM 2016.

in Table 1, but also Feature 2 and 3 since they help to generate correct label sequences. The baseline in Table 4b uses only Feature 1 in Table 3. It can be seen that Feature 8 in Table 1, Feature 2 and 3 in Table 3 lead to decreases of F_1 scores, so they should be excluded[2].

4.3 Final Results

Table 5 shows official evaluation results. The columns "μ-M", "μ-S" and "μ-C" denote microaverages of F1 scores for MWEs (-M), supersenses (-S) and combined (-C), respectively. "Tw", "R" and "TED" indicate F1 scores for tweets, reviews and TED talks, respectively. The results in the last column denote macroaverages of F1 scores across three domains. Compared with other supervised closed condition systems, the performance of our system is not good. This suggests that that there is substantial work to be done on exploring more features to improve our system.

Apart from using simple features, another reason that leads to poor performance of our system is that we use a pipeline model. Since errors generated in the first step are inherited by the second step, the performance of supersense tagging further decreases. Error propagation can be reduced by leveraging joint models. Previous work (Schneider and Smith, 2015) has already leveraged joint models on this issue, so we plan to follow this approach in our future work.

Moreover, it is worth noting that UW-CSE system in the supervised closed condition, achieves competitive results compared with the best scoring systems in the open condition. This suggests that good performance can still be obtained without too much external resources or data.

MWE Features	**Improved F_1**
prefix of w_i	11.2
suffix of w_i	10.4
whether the first character of w_i is uppercase	12.1
whether w_i contains non-alpha or non-numeric characters	7.4
context POS bigram ($t_i t_{i+1}$)	21.4
word + context POS ($t_{i-1} w_i t_{i+1}$)	11.4
Supersense Features	**Improved F_1**
whether swe_i is a noun	2.9
whether swe_i is a verb	5.1
first character and POS combinations of swe_i or mwe_i	6

Table 6: Expanded Feature templates and their improved performance on the development set. w_i denotes the current word and t_i denotes the POS of current word. swe_i or mwe_i denotes the current single or multiple word expression.

4.4 Feature Enrichment

After the evaluation results are released, we expand our feature templates to improve the performance. Table 6 shows the contributions of these expanded features evaluated on the development set and some of them are inspired by (Schneider et al., 2014a). Compared with the performance improvements of MWE detection, the performance of supersense tagging increases less. On the test set, "μ-M", "μ-S" and "μ-C" can achieve 45.6%, 46.1% and 46.0% using all the features proposed in this paper.

4.5 Error Analysis

We calculate false positive (FP) and false negative (FN) errors on the test set. For MWE detection errors in Table 7, a MWE is counted as FP if its boundary is incorrectly identified, and a MWE is counted as FN if it has not been recognized. Table 7 shows that "***_NOUN" is the most difficult pattern to be

[2] In the official evaluation, these features are included due to our coding problems, so this may be one of the reasons due to poor performance.

False Positive		False Negative	
POS Pattern	**Count**	**POS Pattern**	**Count**
NOUN_NOUN	25	NOUN_NOUN	100
VERB_NOUN	20	ADJ_NOUN	59
DET_NOUN	17	VERB_ADP	34
VERB_ADP	16	NOUN_NOUN_NOUN	19
ADP_NOUN	14	PROPN_NOUN	15
ADJ_NOUN	12	DET_NOUN	15
ADJ_ADP	11	VERB_NOUN	14
VERB_PART	10	PROPN_PROPN	13
ADV_ADV	8	ADJ_NOUN_NOUN	11
VERB_ADP_NOUN	7	ADP_NOUN	11

Table 7: Top 10 false positive and false negative patterns for MWE detection.

False Positive		False Negative	
Supersense	**Count**	**Supersense**	**Count**
n.artifact	783	n.person	216
v.stative	629	n.communication	172
n.person	499	n.artifact	168
n.cognition	493	v.change	165
v.social	441	n.act	161
v.cognition	373	v.stative	154
v.communication	303	n.group	136
n.group	269	n.attribute	117
n.communication	253	v.cognition	98
n.time	237	n.location	82

Table 8: Top 10 false positive and false negative supersenses.

recognized. One reason might be that noun phrases often consist of weak MWEs, while weak MWEs are harder to be detected since their vocabularies are more flexible and their meanings are more ambiguous than those of strong MWEs.

For supersense tagging errors in Table 8, a predicted supersense is counted as FP if it is not identical to its corresponding gold supersense, and a gold supersense is counted as FN if it has not been recognized. The single or multiple word expression without a supersense is not taken into account. Table 8 shows that "n.artifact" and "n.person" are more difficult to be tagged. One reason might be "n.artifact" is usually associated with abstract nouns such as "thing" or polysemous words such as "watch". "n.person" recognition might be difficult since various person names lead to many out-of-vocabulary words. This problem may be more serious in the supervised closed condition. In addition, stative verbs are very frequent and also difficult to be disambiguated.

5 Conclusion

We attend the DiMSUM shared task at SemEval 2016 which aims to predict MWEs and supersenses when an English sentence is given. Two submodels, namely CRFs and Maximum Entropy, are explored to detect multiword expressions and supersenses, respectively. Experimental results in the official evaluation suggest that there is substantial work to be done to improve the performance of our system.

In future work, we plan to extend our work in two directions. Firstly, feature templates need to be further expanded and finetuned. Secondly, joint models, which may not only reduce error propagation, but also utilize relations between MWEs and supersenses, can be used to facilitate both subtasks.

References

Giuseppe Attardi, Stefano Dei Rossi, Giulia Di Pietro, Alessandro Lenci, Simonetta Montemagni, and Maria Simi. 2010. A resource and tool for super-sense tagging of italian texts. In *Proceedings of the Seventh conference on LREC*, may.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, August.

Eduard Bejček, Pavel Stranak, and Pavel Pecina. 2013. Syntactic identification of occurrences of multiword expressions in text using a lexicon with dependency structures. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 106–115, June.

Kepa Bengoetxea, Eneko Agirre, Joakim Nivre, Yue Zhang, and Koldo Gojenola. 2014. On wordnet semantic classes and dependency parsing. In *Proceedings of the 52nd Annual Meeting of ACL*.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, T. J. Watson, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18:467–479.

Marie Candito and Matthieu Constant. 2014. Strategies for contiguous multiword expression analysis and dependency parsing. In *Proceedings of the 52nd Annual Meeting of ACL*, pages 743–753, June.

Marine Carpuat and Mona Diab. 2010. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of NAACL*, pages 242–245, June.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy, May.

Massimiliano Ciaramita and Yasemin Altun. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of the 2006 Conference on EMNLP*.

Massimiliano Ciaramita and Mark Johnson. 2003. Supersense tagging of unknown nouns in WordNet. In Michael Collins and Mark Steedman, editors, *Proceedings of the 2003 Conference on EMNLP*.

Matthieu Constant, Anthony Sigogne, and Patrick Watrin. 2012. Discriminative strategies to integrate multiword expression recognition and parsing. In *Proceedings of the 50th Annual Meeting of ACL*.

Dirk Hovy, Anders Johannsen, and Anders Søgaard. 2015. User review sites as a resource for large-scale sociolinguistic studies. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 452–461, New York, NY, USA. ACM.

Anders Johannsen, Dirk Hovy, Héctor Martínez Alonso, Barbara Plank, and Anders Søgaard. 2014. More or less supervised supersense tagging of twitter. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, pages 1–11.

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. 2014. A dependency parser for tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1001–1012, Doha, Qatar, October. Association for Computational Linguistics.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

G. A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41.

Scott Songlin Piao, Paul Rayson, Dawn Archer, and Tony McEnery. 2005. Comparing and combining a semantic tagger and a statistical tool for mwe extraction. *Computer Speech & Language*, 19(4):378–397.

Likun Qiu, Yunfang Wu, and Yanqiu Shao. 2011. Combining contextual and structural information for supersense tagging of chinese unknown words. In *12th International Conference Computational Linguistics and Intelligent Text Processing*.

Carlos Ramisch, Vitor De Araujo, and Aline Villavicencio. 2012. A broad evaluation of techniques for automatic acquisition of multiword expressions. In *Proceedings of ACL 2012 Student Research Workshop*, pages 1–6, July.

Nathan Schneider and Noah A. Smith. 2015. A corpus and model integrating multiword expressions and supersenses. In *Proceedings of the 2015 Conference of the North American Chapter of ACL*, pages 1537–1547, May–June.

Nathan Schneider, Behrang Mohit, Kemal Oflazer, and Noah A. Smith. 2012. Coarse lexical semantic annotation with supersenses: An arabic case study. In *Proceedings of the 50th Annual Meeting of ACL*, pages 253–258, July.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah Smith. 2014a. Discriminative lexical semantic segmentation with gaps: Running the mwe gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206.

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. 2014b. Comprehensive annotation of multiword expressions in a social web corpus. In *Proceedings of LREC'14*, pages 455–461, May.

Nathan Schneider, Dirk Hovy, Anders Johannsen, and Marine Carpuat. 2016. SemEval 2016 Task 10: Detecting Minimal Semantic Units and their Meanings (DiMSUM). In *Proc. of SemEval*, June.

Yutaro Shigeto, Ai Azuma, Sorami Hisamoto, Shuhei Kondo, Tomoya Kouse, Keisuke Sakaguchi, Akifumi Yoshimoto, Frances Yung, and Yuji Matsumoto. 2013. Construction of English MWE dictionary and its application to POS tagging. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 139–144.

Yulia Tsvetkov and Shuly Wintner. 2010. Extraction of multi-word expressions from small parallel corpora. In *Coling 2010: Posters*, pages 1256–1264, August.

Yulia Tsvetkov and Shuly Wintner. 2011. Identification of multi-word expressions by combining multiple linguistic information sources. In *Proceedings of the 2011 Conference on EMNLP*, pages 836–845, July.

Veronika Vincze, Istvn Nagy T., and Jnos Zsibrita. 2013. Learning to detect english and hungarian light verb constructions. *TSLP*, 10(2):6.