Multiword Expressions

IDIOM: Intelligent, Determined Investigators Of MWEs
Brynna Kilcline, Gabby Masini, Ruth Rosenblum, Annika Sparrell
Brandeis University
{brynnakilcline, gmasini, ruthros, asparrell}@brandeis.edu

1. Introduction

Our goal is to identify and categorize all multiword expressions (MWEs) in our data. The annotated data will become training data for an ML model that will be able to identify different types of multiword expressions in text, which could be useful for downstream tasks such as machine translation, translation detection, synonym detection, and syntactic parsing.

The annotation task comprises finding multiword expressions and labeling their type, which we frame as spans. We define a multiword expression as a string of multiple words which act as a single semantic or syntactic element.

1.1. Related Work

Katz and Giesbrecht (2006) show that it is possible to use local context to distinguish compositional and non-compositional MWEs by measuring the similarity (1) between matching compositional and non-compositional expressions and (2) between the MWEs and their component words using the cosine of LSA vectors.

In "Multiword Expressions: A Pain in the Neck for NLP", Sag and others (2002) discuss the merits and pitfalls of using statistical vs rule-based techniques for identifying multiword expressions in the newly growing field. The authors boldly state that a purely statistical approach is ineffective, and rule-based techniques are a necessary addition into any method for identifying MWEs.

In "An empirical model of multiword expression decomposability", Baldwin et al. (2003) attempt to define and predict the compositionality of MWEs – that is, to predict the extent to which the meaning of an MWE can be derived from the meaning of its individual components. They show that MWEs that are more decomposable are likely to have more hyponyms – especially for low-frequency noun-noun compounds and verb-particle constructions. This is relevant for our project, since part of formulating our annotation task involved defining what to consider an MWE, and this definition relies in part of the semantic decomposability of the phrases in question.

2. Guidelines

The specification includes five types of multiword expressions–idioms, noun-noun compounds, verb-particle constructions, light verbs, and other–based on our background linguistic knowledge and preliminary research. We skimmed the data for examples of each type of MWE to include in the first draft of the guidelines. Using those initial guidelines, each team member individually annotated some data. We then went through our annotations together and the notes we had made and clarified the guidelines to reflect our discussion. The most significant change was to the definition of noun-noun compounds. We originally defined multiword expressions as non-compositional phrases, but we determined that this semantic definition was too strict in the case of noun-noun compounds. This led to changing the syntactic definition for this category. Instead, we determined that a noun-noun compound is an expression formed of a sequence of two or more nouns that come together to form what syntactically behaves as a single noun.

Halfway into the internal annotation process, we switched from using attribute to span labels to mark the type of MWE in order to simplify the annotators' task and post-processing. We simplified the guidelines slightly for the crowd-sourcing workers by collapsing the *other* category into the *idiom* category, adding simplified definitions of each tag type, and giving more examples of what a multiword expression is and is not.

3. Data

3.1. Original Dataset

The Corpus of Contemporary American English (Davies, 2008) contains 950 million words of data in American English from 1990 to 2019 in various genres, including spoken conversations, news and magazines, and literature (fiction). The data is tokenized and tagged for lemmas and parts of speech. Access to the full corpus requires paid registration, but we were able to download a free sample of about 1/100th of the corpus (8.9 million words).

3.2. Preprocessing

We took the plain text data of the sample corpus and ran the SpaCy sentence segmentizer on the data to obtain 10-sentence chunks of the text which we could present to an annotator. We then ran the SpaCy part of speech tagger on the data in order to gain some information on potential MWEs that are based on syntactic category, such as noun-noun compounds and verb-particle constructions. Finally, we ran through a lexical lists of light verb and idiom lemmas to highlight potential candidates for those categories as well. Another step taken in preprocessing was replacing sensitive information that was originally present in the data as a span of @-symbols with spaces in order for the tagger to function properly.

We sampled several hundred documents (10-sentence chunks) from the free COCA sample. The documents were shuffled to ensure a representative balance of the different documents and genres. Documents were then selected by iterating through the data until a minimum number of

candidates for each type of MWE was reached. A candidate MWE had a syntactic construction, lemma, or phrase matching the constraints highlighted above. We assigned each document an ID and shuffled them before giving them to our annotators.

3.3. Final Dataset

The group completed our annotations in Label Studio and exported them as a JSON file. In order to compute inter-annotator agreement metrics and conduct machine learning processes, we converted the JSON output to BIO format using the label names and token offsets.

Type	Amount
Noun-Noun Compound	749
Verb-Particle Construction	288
Light Verb Construction	117
Idiom	280
Other	49
All MWEs	1483

Table 1: Number of instances of each type of MWE identified in the data.

Overall, we had 347 total labeled documents, and each document was approximately 10 sentences long. Table 1 demonstrates the overall amount of labels that we had of each type in our corpus. Noun-noun compounds were the most common, followed by verb-particle constructions and idioms. The Other category was the least common.

4. Inter-annotator Agreement and Adjudication

The annotators were eight graduate students who had taken at least one syntax and one semantics course, including ourselves. Each person annotated between 12 and 113 documents, the mean being 70 documents per annotator. One of our team member's data were not used in the end due to a discrepancy in the formatting. Although we set up crowdsourcing in Scale, we encountered errors with the set-up of the initial and review phase tasks and decided the quality of the crowd-sourced data was not sufficient to warrant addressing the errors. Thus, we ended up with data from seven annotators.

117 of the 374 documents were annotated by more than one person. Of these, some had two annotators and some had three. We computed inter-annotator token-level agreement over these documents using Fleiss's Kappa, shown in Table 2. According to the scale specified by Landis and Koch (1977), the overall agreement was moderate. Agreement for noun-noun compounds was substantial, verb-particle construction agreement was fair, and the others were slight.

Two of the team members performed adjudication on the documents with more than one annotator. The adjudicators kept annotations that were agreed upon between at least two annotators and resolved disagreeing annotations using their best judgment.

Type	IAA Score
Noun-Noun Compound	0.7878
Verb-Particle Construction	0.3500
Light Verb Construction	0.1855
Idiom	0.1651
Other	0.1034
All MWEs	0.5793

Table 2: Token-level agreement for each type of MWE calculated using Fleiss's Kappa.

	Precision	Recall	F1	Support
I	0.61	0.06	0.11	374
O	0.94	1.00	0.97	5930
Macro Average	0.77	0.53	0.54	6304
Weighted Average	0.92	0.94	0.92	6304
Accuracy		0.94		6304

Table 3: Test set results for span labeling using logistic regression.

5. Machine Learning Baseline

We ran two models to test how our dataset would perform for a machine learning task, logistic regression and CRF.

For logistic regression, we tried two styles of task, both an NER-style (span labeling) task to see how well the computer could distinguish between tokens that belonged to a MWE and tokens that did not, as well as a classification task in which the model also had to identify which type of MWE a token belonged to. For the logistic regression model's features, we used scikit-learn's DictVectorizer tool to turn a dataframe of individual tokens paired with their appropriate class into features for a linear model. For the span labeling task, we collapsed all MWE categories into the label I to represent a positive instance of an MWE, while for the classification task we used the original categories of I-IDIOM, I-LIGHT_V, I-NN_COMP, and I-V-P_CONSTRUCTION from our postprocessed data, although we collapsed OTHER into IDIOM. For logistic regression, we tuned the hyperparameter C in the dev set, with values of 1.0 and 0.5, and found that it performed slightly better (0.002 points more accurate) with C=1.0, which is the default of the model. For the span labeling task, we obtained an accuracy of 0.95 on the development set and an accuracy of 0.94 on the test set. Full test set results are given in Table 3. For the classification task, we also obtained an accuracy of 0.95 on the development set and 0.94 on the test set. Full test set results are given in Table 4.

We also attempted a CRF model (using sklearn-crfsuite) in order to try to better classify sequences. Each document was treated as a sequence where each word was tagged with our aforementioned categories (I-IDIOM, I-LIGHT_V, I-NN_COMP, and I-V-P_CONSTRUCTION, and O). We tuned the hyperparameters c1 and c2 on the dev set with these being the coefficients of l1 and l2 regularization respectively. For both, combinations of the hyperparameters 0.05, 0.1, 0.15, 0.2, and 0.3 were tested. Ultimately, the

	Precision	Recall	F1	Support
I-IDIOM	1.00	0.00	0.00	103
I-LIGHT_V	1.00	0.06	0.11	33
I-NN_COMP	0.00	0.00	0.00	153
I-V-P_CONSTRUCTION	0.79	0.15	0.26	71
O	0.94	1.00	0.97	5944
Macro Average	0.75	0.24	0.27	6304
Weighted Average	0.92	0.94	0.92	6304
Accuracy	0.94			6304

Table 4: Test set results for classification using logistic regression.

best combination was c1=0.1 and c2=0.5. This gave an F1 score of 0.9349 on the dev set. With these hyperparameters, we were able to achieve an accuracy of 0.95 and a macroaveraged F1 score of 0.44. More detailed results are provided in 5.

Ultimately, logistic regression was not a very good model to use for this task, as it was a token level classification task with a large proportion of the tokens belonging to a single class (O - not a multiword expression). The CRF model did do better than the logistic regression model, but it similarly did not perform particularly well. However, both models did perform better than chance, as a classifier that predicts O every time would only obtain an accuracy of 0.855, significantly lower than the actual performance. In the future, we could try changing the features to represent entire spans rather than single tokens, or classify using other models.

6. Future directions

For simpler annotations and likely improved agreement, we would collapse the *idiom* and *other* category like we did for the crowd sourcing task setup. Furthermore, if we wanted to use crowd sourcing and hopefully get useful results, we could attempt to make the task easier and the instructions shorter and require less linguistic knowledge. For example, we could potentially set up different tasks for each of our MWE labels (i.e. one group would annotate noun-noun compounds, another verb-particle constructions, etc.). Additionally, we could perform other types of IAA metrics, such as Cohen's Kappa, on the data that we have exactly two annotators for. We could also run our dataset through other machine learning models, such as other neural models.

7. Acknowledgements

Thank you to Group 3 for annotating for us as our partner group.

8. Bibliographical References

Baldwin, T., Bannard, C., Tanaka, T., and Widdows, D. (2003). An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 89–96, Sapporo, Japan, July. Association for Computational Linguistics.

Katz, G. and Giesbrecht, E. (2006). Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19.

Landis, J. R. and Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, pages 363–374.

Sag, I. et al. (2002). Multiword expressions: A pain in the neck for NLP. In *Lecture Notes in Computer Science*, volume 2276, pages 1–15.

9. Language Resource References

Davies, M. (2008-). Corpus of Contemporary American English (COCA).

	Precision	**Recall**	**F1**	**Support**
I-IDIOM	0.27	0.06	0.10	96
I-LIGHT_V	1.00	0.17	0.30	40
I-NN_COMP	0.68	0.22	0.34	134
I-V-P_CONSTRUCTION	0.65	0.38	0.48	63
O	0.96	0.99	0.97	5652
Macro Average	0.71	0.37	0.44	5985
Weighted Average	0.94	0.95	0.94	5985
Accuracy	0.95			5985

Table 5: Test set results for classification using a CRF model.