Towards a Dataset for Evaluating Multiword Predicate Interpretation in Context

Report from Nathan Schneider's Short-Term Scientific Mission to Israel under PARSEME (Dec. 19–Jan. 6)

Nathan Schneider, University of Edinburgh
Omri Abend, Hebrew University of Jerusalem

January 31, 2016

Abstract

Many multiword expressions (in the sense of Baldwin and Kim, 2010) are verbal predicates taking one or more arguments—including light verb constructions, other verb-noun constructions, verb-particle constructions, and prepositional verbs. We frame a lexical interpretation task for such multiword predicates (MWPs): given a sentence containing an MWP, the task is to predict other predicates (single or multiword) that are entailed. Preliminary steps have been taken toward developing an evaluation dataset via crowdsourcing.

Natural languages have large vocabularies, especially taking into account multiword expressions (MWEs; Baldwin and Kim, 2010), which are so numerous in English that they cannot be listed exhaustively by traditional lexicographic methods. Computational lexicons such as WordNet (Fellbaum, 1998) are known to be limited in their coverage of MWEs. But many MWEs are sufficiently frequent in text that it is imperative for information extraction and natural language understanding systems to process them. One aspect of this is recognizing how an individual MWE is related to other lexical expressions, such as through synonymy or entailment.

Automatic, data-driven techniques can help us in assembling broad-coverage knowledge about lexical relationships. For example, distributional methods over large corpora can in principle be used to extract graphs of lexical entailments, such as 'PERSON *buy* ARTIFACT ⇔ PERSON *make a purchase of* ARTIFACT ⇒ PERSON *take owernship of* ARTIFACT ⇒ PERSON *own* ARTIFACT ⇔ ARTIFACT *belong to* PERSON' (Berant et al., 2011, 2012, 2015; Abend et al., 2014). This sort of distributional learning, which involves minimal supervision, is potentially a powerful way to induce semantic lexicons for applications such as question answering. Currently, such graphs can be evaluated by looking for overlap with Zeichner et al.'s (2012) gold-standard dataset of positive and negative entailment pairs. Loukou (2016) has successfully generalized the distributional learning methods to light verb constructions with 2 (typed) arguments, showing a small improvement on the Zeichner et al. (2012) dataset. Unfortunately, that dataset is too small for a robust evaluation.

We propose methods for creating a much larger **gold-standard dataset** recording judgments of semantic entailment relations among English predicates. Each item will involve (i) a **premise sentence** containing a multiword predicate like *take ownership of*—the **target predicate**; (ii) single- or multiword predicates that may be related to the target—the **entailment candidates**; and (iii) human judgments of whether each candidate is or is not entailed by the premise sentence.

1 Sentence-to-predicate entailment task

We propose what is (to the best of our knowledge) a novel framing of an entailment task as pairing a full *sentence* with candidate entailed *predicates*.

1.1 Motivation and setup

The Recognizing Textual Entailment (RTE) challenge, in its canonical form, requires a system to predict whether a natural language **hypothesis** sentence logically follows from a **premise** sentence or passage (Dagan et al., 2013). Because humans can draw on extensive knowledge of both language and the world in comprehending sentences, they can recognize entailments that are extremely challenging for systems. In its most general form, an entailment task may rely on any type of knowledge, rendering it "AI-complete". For example, the following might reasonably be considered a true entailment pair:

(1) a. *(Premise)* Until they ground to a halt, Maxine had failed to notice that the fuel gauge was pointing on "EMPTY".

b. *(Hypothesis)* The vehicle was out of gas.

Recognizing the entailment relationship in (1) requires lexical knowledge (*gas* being an alternative word for *fuel* in this context), but also the ability to recognize that the premise invokes a scene that takes place in a motor vehicle, and the capability to make temporal and causal inferences.

Rather than try to solve this thorny problem at once, it is reasonable to focus on subproblems. The line of work noted above narrows the scope of the problem to recognizing **lexical entailments**. The dataset of Zeichner et al. (2012) consists of two-argument predicates where the hypothesis is a simplified sentence (extracted from the web by ReVerb), and the premise was artificially generated by substituting a distributionally similar predicate.[1] As an alternative to artificially generating paraphrases, other corpora (such as NewsSpike; Zhang and Weld, 2013) were sampled by mining several news stories about the same event, and extracting sentences that convey similar information.

We consider a third alternative that avoids artificial paraphrases without limiting the data to events reported multiple times: namely, we propose that the premise should be a full, naturally-occurring sentence, while the hypotheses should be isolated predicates (possibly with argument slots filled in with vague pronouns: *somebody swam*, *somebody purchased something*). The premise sentence establishes the context to avoid the confound of word sense ambiguity, while the hypothesis encourages the inference to be lexical in nature. Our setup would not force the hypothesis to be an entailment specifically of the target predicate in the premise; it may be entailed by the sentence as a whole (e.g., *She finished her food* ⇒ *somebody ate*). Still, we expect that if candidates are generated based on the target predicate, most of the true entailments will follow from that predicate.

As a starting point, our focus will be on premise sentences likely to contain a predicate that is a **light verb construction** (e.g., *make a decision*; *make sure*; *take advantage of*; *pay attention to*) or other idiomatic **verb-noun combination** (*come to blows*; *kick the bucket*).

1.2 Proposed crowdsourcing task

Inspired in part by the recent **SNLI** dataset by Bowman et al. (2015), which contains human-authored full-sentence entailments (often simplifications)[2] of image captions, we propose to elicit semantic judgments by crowdsourcing. If pilot tests with local annotators are successful, we will launch a task on Amazon

[1] As "predicates", ReVerb's shallow heuristics extract words or phrases linking two entities—if longer than one word, it may be an MWE, or a compositional phrase like "could be exchanged for" (Abend et al., 2014).

[2] It also contains contradictions, as well as statements that are neither entailed nor contradicted.

Mechanical Turk with the goal of collecting judgments for 10,000 pairs. The resulting dataset will be released for the benefit of MWE and textual entailment research.

A tentative proposal for the task to be completed by crowd workers is as follows:

> Imagine a scene where the following sentence applies: [premise sentence, e.g.] *Two females (in blue and bright orange shirt respectively) taking a stroll*
>
> In your imagined scene, is it true that someone or something is:
>
> 1. getting dressed
> 2. smiling
> 3. eating
> 4. walking
> 5. photographing someone/something
> 6. posing
> 7. talking
>
> Check all that are true in your imagination of the scene. Name two more actions that are also happening:
>
> 1. _____
> 2. _____
>
> Let us know if you do not understand the prompt or any of the options.

The second half of the task can elicit responses to be offered as possible candidates to subsequent participants. For each premise sentence, we would solicit responses from at least 7 participants, which would help us to filter out noise.

Note that we have not highlighted the target multiword predicate itself, as we feel this would complicate the instructions. Thus, some of the responses may indicate entailments of other parts of the sentence, unrelated to the multiword predicate. We can mitigate this by instantiating the task for several different sentences containing the target predicate, and intersecting the responses.

1.3 Pilot dataset

We created a pilot dataset of sentences including a multiword predicate, drawing on two sources:

- **STREUSLE:** The STREUSLE corpus (Schneider et al., 2014b) contains comprehensive gold-standard annotations of MWEs in online reviews. We examined STREUSLE 3.0 sentences with gold verb+noun annotations and selected 67 of them. For example, **gain entry**: Three weeks ago , burglars tried to gain_entry into the rear of my home .
- **SNLI:** We ran the AMALGrAM tagger (Schneider et al., 2014a) on caption sentences (limited to premises with a hypothesis deemed a true entailment by 4 or 5 annotators, and subject to a word length filter). We examined instances containing a verb+noun automatically tagged as an MWE, and selected 52 of them, retaining the corresponding hypothesis sentence. For example, **give thumbs**

up: *(Premise)* A race_car_driver smiles and gives_ the _thumbs_up before a race . *(Hypothesis)* a race_car_driver is about_to_race

Most predicates have only one instance, but a handful are duplicated. The unique predicates are listed below in Appendix A.

2 Obtaining entailed predicate candidates

We investigated several methods to obtain, given a multiword predicate, candidate entailed predicates. To construct an evaluation corpus, it is to our benefit to use several methods of obtaining candidates, to reduce the possibility that entailment systems being evaluated would be artificially advantaged or disadvantaged by using any particular method.

2.1 Mining predicates linking frequently coocurring named entities in the NYT corpus

Taking inspiration from the distributional entailment graph literature discussed above, we considered similar techniques to extract clusters of predicates (including multiword predicates) that relate the same pair of entities in the *New York Times* corpus. We ran EasySRL, a CCG-based semantic role labeler (Lewis et al., 2015) to preprocess sentences in several months' worth of articles. Using capitalization as a rough indicator for named entities,[3] we listed capitalized core arguments often appearing together as core arguments of the same predicate.

The qualitative results, however, were disappointing. This technique only worked for a few pairs of entities (individuals, countries, companies, sports teams, political organizations) sufficiently engaged with one another that the NYT wrote many stories about their relations. E.g., nations at war; the president and Congress; companies in a high-profile competition or acquisition. A lot of the relationships involved predicates of communication where there was a content clause describing the topic of discussion; without that, the pair of entities and the predicate are not very informative.

Because this strategy was not producing all that many interesting clusters of related predicates, we decided to abandon it in favor of other techniques.

2.2 Existing lexicons

Some existing computational lexicons contain light verb constructions that are semantically related to other entries. We examined WordNet (Fellbaum, 1998), with hopes of exploiting synset groupings and relations between synsets, and FrameNet (Fillmore and Baker, 2009), with hopes of exploiting frame groupings and relations between frames.

Unfortunately, both resources had poor coverage of the 111 multiword predicates we selected for our pilot. FrameNet's list of supports[4] (light verbs being a type of support) covers only about 20 (though we did not systematically normalize the inclusion of determiners, prepositions, etc. when comparing the two lists). Based on manually checking a sample, WordNet's coverage is probably lower still.

We therefore have concluded that WordNet and FrameNet are not presently very useful for generating candidate entailed predicates. However, we will investigate the new version of PropBank (Kingsbury and

[3] We filtered out pronouns, articles, and other function words often appearing (capitalized) at the beginning of a sentence.

[4] http://www1.icsi.berkeley.edu/~warrenmc/mwe_supps.txt, with 1613 total entries

Palmer, 2002) when it is released, as it annotates light verbs (Bonial et al., 2014) and groups together morphologically related predicates across parts of speech.

2.3 PPDB

The Paraphrase Database (PPDB; Ganitkevitch et al., 2013; Pavlick et al., 2015b) contains a large number of phrase pairs automatically extracted from corpora. Version 2.0 contains crowdsourced entailment annotations for each pair (Pavlick et al., 2015a). For example, the pairs whose source phrase starts with *make* include:

- make a commitment ||| commit (ForwardEntailment)
- make a reply ||| responded (ForwardEntailment)
- make further progress ||| advance (Equivalence)
- make reparation ||| compensate (OtherRelated)

From the XXL-sized release of PPDB, we managed to extract 2228 unique multiword phrases with *take*, *make*, *do*, *get*, *have*, *pay*, or *give* that have an entailing, entailed, or equivalent pair. This includes verb-particle constructions as well as light verb constructions.

2.4 Vector space models

Srivastava and Hovy (2014) proposed a vector space model that can be used to measure distributional similarity between single words and "motifs", multiword phrases including multiword expressions. It thus may be possible to query their model with a multiword predicate to retrieve semantically similar candidates. (In fact, our crowdsourcing task could be viewed as a way of evaluating such models: if the model's similarity scores are informative, one hypothesis would be that a greater number of highly similar pairs would lead to an entailment judgment than less similar pairs.)

3 Future work

The next step will be to iterate on the pilot task by giving it to local participants, then publish it on Amazon Mechanical Turk to collect judgments for a large number of multiword predicates. We also anticipate varying the task instructions to elicit different subtypes of entailment relations, e.g., temporal, causal, and hypernymy relations (or else creating a separate task to classify entailments under one of these subtypes).

Acknowledgments

For this research, the first author was supported by travel funding awarded under the COST program, and an EU IST Cognitive Systems IP EC-FP7-270273 grant, "Xperience", awarded to Mark Steedman. We thank Mark Steedman and Felisia Loukou for establishing the need for this research, and Vered Shwartz for a useful suggestion regarding PPDB.

A Pilot multiword predicates

bite the dust
blow bubble
catch my eye
change lightbulb
change mind
come over budget
cut deal
cut price
fix dog
fill role
gain entry
get act together
get chance
get rid of glare
give a call
give a chance
give a darn
give a try
give deal
give impression
give it shot
give ride
give second chance
give the finger
give thumb up
go out of business
go the extra mile
have a clue
have a good time
have a laugh
have a problem
have come a long way
have complaint
have experience
have fun
have hair cut
have problem
have surgery
hit the ground
hit the nail on the head
jaw drop
jimmy rig
keep company
keep in mind
kiss *ss
know stuff
lead the way
make a buck
make appointment
make catch
make decision
make exchange
make face
make mistake
make peace sign
make purchase
make recommendation for
make repair
make run
make sale
make the drive
make toast
make way
pass time
pay attention
pay attention to
picture take
ply trade
rest eyes
return favor
shake hand
spend time
spread the word
stand the test of time
steal a base
step it up notch
strike pose
take a bath
take a bite
take a break
take a crack
take a drag on
take a look
take a moment
take a nap
take a photograph
take a picture
take a rest
take a risk
take a shot
take a stroll
take a turn
take a walk
take advantage
take care of
take down number
take name
take note
take order
take part
take place
take pride
take the time
take time
take turn
tell story
tell the truth
throw a tantrum
throw birthday party
treat like dirt
trust gut
turn the corner
waste time

References

Omri Abend, Shay B. Cohen, and Mark Steedman. Lexical inference over multi-word predicates: a distributional approach. In *Proc. of ACL*, pages 644–654, Baltimore, Maryland, USA, June 2014.

Timothy Baldwin and Su Nam Kim. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*, pages 267–292. CRC Press, Taylor and Francis Group, Boca Raton, FL, 2010. ISBN 978-1420085921.

Jonathan Berant, Ido Dagan, and Jacob Goldberger. Global learning of typed entailment rules. In *Proc. of ACL-HLT*, pages 610–619, Portland, Oregon, USA, June 2011.

Jonathan Berant, Ido Dagan, and Jacob Goldberger. Learning entailment relations by global graph structure optimization. *Computational Linguistics*, 38(1):73–111, March 2012.

Jonathan Berant, Noga Alon, Ido Dagan, and Jacob Goldberger. Efficient global learning of entailment graphs. *Computational Linguistics*, 41(2):221–263, April 2015.

Claire Bonial, Julia Bonn, Kathryn Conger, Jena D. Hwang, and Martha Palmer. PropBank: semantics of new predicate types. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proc. of LREC*, pages 3013–3019, Reykjavík, Iceland, May 2014.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proc. of EMNLP*, pages 632–642, Lisbon, Portugal, September 2015.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. *Recognizing Textual Entailment: Models and Applications*. Number 23 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool, San Rafael, CA, July 2013.

Christiane Fellbaum, editor. *WordNet: an electronic lexical database*. MIT Press, Cambridge, MA, 1998.

Charles J. Fillmore and Collin Baker. A frames approach to semantic analysis. In Bernd Heine and Heiko Narrog, editors, *The Oxford Handbook of Linguistic Analysis*, pages 791–816. Oxford University Press, Oxford, UK, December 2009.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. PPDB: The Paraphrase Database. In *Proc. of NAACL-HLT*, pages 758–764, Atlanta, Georgia, USA, June 2013.

Paul Kingsbury and Martha Palmer. From TreeBank to PropBank. In *Proc. of LREC*, pages 1989–1993, Las Palmas, Canary Islands, May 2002.

Mike Lewis, Luheng He, and Luke Zettlemoyer. Joint A* CCG parsing and semantic role labelling. In *Proc. of EMNLP*, pages 1444–1454, Lisbon, Portugal, September 2015.

Felisia Loukou. *Light verb constructions in distributional entailment graphs*. MSc thesis, University of Edinburgh, Edinburgh, Scotland, UK, 2016.

Ellie Pavlick, Johan Bos, Malvina Nissim, Charley Beller, Benjamin Van Durme, and Chris Callison-Burch. Adding semantics to data-driven paraphrasing. In *Proc. of ACL-IJCNLP*, pages 1512–1522, Beijing, China, July 2015a.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proc. of ACL-IJCNLP*, pages 425–430, Beijing, China, July 2015b.

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

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. Comprehensive annotation of multiword expressions in a social web corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard,

Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proc. of LREC*, pages 455–461, Reykjavík, Iceland, May 2014b.

Shashank Srivastava and Eduard Hovy. Vector space semantics with frequency-driven motifs. In *Proc. of ACL*, pages 634–643, Baltimore, Maryland, USA, June 2014.

Naomi Zeichner, Jonathan Berant, and Ido Dagan. Crowdsourcing inference-rule evaluation. In *Proc. of ACL*, pages 156–160, Jeju Island, Korea, July 2012.

Congle Zhang and Daniel S. Weld. Harvesting parallel news streams to generate paraphrases of event relations. In *Proc. of EMNLP*, pages 1776–1786, Seattle, Washington, USA, October 2013.