Exploiting Open IE for Deriving Multiple Premises Entailment Corpus

Martin Víta
NLP Centre, Faculty of Informatics
Masaryk University
Botanická 68a, 602 00 Brno
Czech Republic
info@martinvita.eu

Jakub Klímek
Department of Software Engineering
Faculty of Mathematics and Physics
Charles University
Malostranské nám. 2/25, 118 00 Prague 1
Czech Republic
klimek@ksi.mff.cuni.cz

Abstract

Natural language inference (NLI) is a key part of natural language understanding. The NLI task is defined as a decision problem whether a given sentence – hypothesis – can be inferred from a given text. Typically, we deal with a text consisting of just a single premise/single sentence, which is called a single premise entailment (SPE) task. Recently, a derived task of NLI from multiple premises (MPE) was introduced together with the first annotated corpus and corresponding several strong baselines. Nevertheless, the further development in MPE field requires accessibility of huge amounts of annotated data. In this paper we introduce a novel method for rapid deriving of MPE corpora from an existing NLI (SPE) annotated data that does not require any additional annotation work. This proposed approach is based on using an open information extraction system. We demonstrate the application of the method on a well known SNLI corpus. Over the obtained corpus, we provide the first evaluations as well as we state a strong baseline.

1 Introduction

Natural language inference (NLI), formerly known as recognizing textual entailment (RTE) task – as a part of natural language understanding (NLU) – belongs to one of the most prominent problems in NLP for more than ten years. Generally, the NLI task is to classify the relationship between a given text and a given hypothesis: whether the hypothesis can be inferred from the text. The task is typically formulated in a "sentence-pair setting", i. e., the text is just a single sentence. According to (Lai et al., 2017), we refer this setting as a *single premise entailment* (SPE for short). Current state-of-the-art approaches are based on deep learning and/or ensemble methods. Over the years, solid resources for supervised learning for SPE were developed. In contrast, problems related to NLI and/or problems derived from NLI, like relation inference (Levy and Dagan, 2016), question entailment (Abacha and Demner-Fushman, 2016), partial/facet entailment (Levy et al., 2013) and (Nielsen et al., 2009), and others, are strongly under-resourced. For further development in these fields, this fact may be limiting.

Recently, a task of NLI from multiple premises was proposed in (Lai et al., 2017) – the idea is based on relaxing the common assumption that the premise is just a single sentence. Again, according to this paper, we will call this derived of NLI task *multiple premises entailment* (MPE for short). Similarly to other mentioned entailment tasks, MPE is also under-resourced: to best of our knowledge, there exists only one annotated corpus (introduced in the original paper) for MPE.

The main aim of this work is to describe a novel method of preparing MPE annotated corpora from existing NLI SPE ones. It is based on using open information extraction systems and on several plausible assumptions. Then we apply the proposed method on a concrete corpus and provide the first evaluation and we state a strong baseline for MPE task on this obtained corpus.

2 Preliminaries and Related Work

In this section we are going to put MPE task in context, describe briefly the notion of open IE and recall two entailment tasks where textual tuples play a certain role.

Proceedings of Recent Advances in Natural Language Processing, pages 1257–1264,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_144

2.1 NLI Task and Notable Corpora for NLI

There exist several definitions of NLI or, formerly, RTE task. Indeed, the differences among them are rather subtle and have no real consequences for NLP. For completeness, we provide the original definition from (Dagan et al., 2005): "*We say that T entails H if humans reading T would typically infer that H is most likely true.*" The deep insight into the nature of NLI from the logical and philosophical point of view is provided in (Korman et al., 2018).

Originally, RTE was proposed as a binary decision task (entailment/non-entailment). Later, the 3-way task (entailment/neutral/contradiction) became more frequent.

Nowadays, there exists a number of annotated corpora for the NLI task. At the beginning of RTE investigations, it was a collection of RTE corpora created for Pascal/NIST/SemEval challenges. A comprehensive overview of older RTE corpora is provided in (Bentivogli et al., 2017).

A massive development in the field of NLI using deep learning approaches was started after the release of the Stanford NLI corpus (Bowman et al., 2015), probably the most widely used annotated corpus for NLI, containing approx. 570K of annotated sentence text-hypothesis pairs. This corpus was later followed by MultiGenre NLI corpus – MultiNLI (Williams et al., 2018) – of a comparable size, but with a wider range of genres, including spoken language, newspapers, 9/11 etc. Both of these corpora were constructed in a similar way: given a sentence/premise, the annotators were asked to write a sentence that is entailed by the premise, a sentence that is contradictory to the premise, and a sentence neutral w. r. t. the premise, i. e., such that its truth value is independent to the truth value of the premise. According to the classification presented in the paper (Poliak et al., 2018), these two corpora belong to the *human elicited* category. The paper also provides a comprehensive analyses of SNLI, MultiNLI as well as an overview of more recent and specific NLI corpora.

2.2 Natural Language Inference from Multiple Premises

As already mentioned, the novel NLI task that is based on inference over multiple premises was recently introduced in (Lai et al., 2017). Given four premise sentences and one hypothesis sentence, the task is to label this premises-hypothesis pair in a standard 3-way manner – entailment, neutral, or contradiction.

This work was inspired by the *Approximate entailment task* (Young et al., 2014), that arises from processing the image captions – the task is to decide whether a brief caption h (the hypothesis) can describe the same image as a set of captions $P = \{P_1, \ldots, P_N\}$ known to describe the same image (the premises).

The (only one) MPE corpus[1] introduced in the paper (Lai et al., 2017) was created upon the FLICKR30K dataset (Plummer et al., 2015). Hypotheses were generated in by simplifying either a fifth caption describing the same image or a caption corresponding to a different image and given the standard 3-way tags (Poliak et al., 2018). The simplification process relies on the denotation graph (Young et al., 2014) – it is based on normalization and reduction rules (e. g. lemmatization, dropping modifiers and prepositional phrases, replacing nouns with their hypernyms, extracting noun phrases), see (Lai et al., 2017). Each hypothesis has at most 50% overlap with the words in its corresponding premises. The MPE corpus contains 8000 items in the training set, 1000 items in the development set and 1000 in the test set.

To provide a better idea about the corpus, here is an example of positive (entailment) item taken again from (Lai et al., 2017):

Premises:

1. *Two girls sitting down and looking at a book.*
2. *A couple laughs together as they read a book on a train.*
3. *Two travelers on a train or bus reading a book together.*
4. *A woman wearing glasses and a brown beanie next to a girl with long brown hair holding a book.*

Hypothesis: *Women smiling.*
Label: ⇒ ENTAILMENT

In the paper, the authors also investigate the relation between MPE and the standard (SPE) entailment. In this particular MPE task/corpus, each premise consists of four independently written sentences and, using crowdsourcing, single-premise entailment labels for each individual

[1] https://github.com/aylai/MultiPremiseEntailment/tree/master/data/MPE

single-premise-hypothesis pair in the DEV dataset were obtained. Based on these individual labels, it has been shown that majority voting strategies as well as more sophisticated rule based approaches over single labels to obtain the final MPE label do not lead to sufficient results, hence MPE cannot be trivially reduced to multiple SPE tasks.

2.3 Information Extraction (IE) and Open Information Extraction (Open IE)

Information extraction is generally a process of transforming an unstructured textual information into a structured representation in the form of relational phrase and its arguments, usually `(arg1 ; rel-phrase; arg2)`, see (Niklaus et al., 2018). Information extraction deals with a predefined relation vocabulary.

In contrast, in *open information extraction* introduced by (Banko et al., 2007), this assumption is relaxed, i. e., we do not require a fixed vocabulary of relations. Open information extraction systems extract textual n-tuples that represent basic propositions asserted by a sentence (Stanovsky et al., 2018). An example of a result of open IE process taken again from (Stanovsky et al., 2018):

INPUT: *Mercury filling, particularly prevalent in the USA, was banned in the EU, partly because it causes antibiotic resistance.*

OUTPUT:

- `(mercury filling; particularly prevalent; in the USA)`
- `(mercury filling; causes; antibiotic resistance)`
- `(mercury filling; was banned; in the EU; partly because it causes antibiotic resistance)`

2.4 Relational Entailment/Relation Inference

Open information extraction over particular parts of an NLI corpus (hypotheses) was already exploited in (Víta, 2018) in order to obtain "sentence-textual tuple" entailment pairs when introducing a task of *relational entailment*. This task can be employed for checking facts in open knowledge bases, i. e., sets of extracted tuples, see (Mausam, 2016).

Another entailment task based on relational tuples, was introduced by Levy et al. in (Levy and Dagan, 2016) together with a new annotation method for collecting data (on relation inference) in context: the inference task was transformed into simple factoid question answering. The resulting annotated corpus has a form of "textual triple-textual triple plus entailment label".

Indeed, in both cases, textual tuples are a "subject of entailment" – in the final corpus, the tuples are inputs for the entailment decision. In this paper we are going exploit open IE in a different way – to create certain sentences.

3 Methods

In this section we are going to describe a general method of constructing an MPE corpus from a given single premise entailment corpus using an open IE system and also its evaluation. Concrete implementation details are provided in the next section.

3.1 Creating MPE Corpus from SPE One

The main idea of this proposed approach is based on the following observations:

- From longer sentences, it is usually possible to extract multiple textual n-tuples.
- Results of open IE systems is naturally interpretable when reading from left to right (Stanovsky et al., 2018), hence they correspond with sentences.
- The entailment label in an SPE task can be used even for tasks where premise is represented in a "semantically equivalent form".

In order to provide a compact notation, we introduce the following convention: let us denote a set of word types contained in a sentence or a textual n-tuple t by a symbol $||t||$. Let $e(s)$ be a set of textual n-tuples extracted by an open IE system from a sentence s and, finally, let $s(t)$ be a string obtained from a textual n-tuple t by removing auxiliary symbols (brackets and semicolons) – this refers to the second observation: we assume that $s(t)$ can be treated as a sentence – it is a subject of further investigations.

Given a premise P and a hypothesis H and a corresponding entailment label L, we transform this P-H pair into a set of multiple premises $M(P) = \{s(t) \mid t \in e(P)\}$ accompanied with the unchanged hypothesis H a s well as unchanged label L, iff the following conditions hold:

1. $|e(P)| > 1$, i. e., we are interested only in such cases, when more than one tuple is extracted from a given premise,

2. $(\cup_{t \in e(P)} ||t||) \setminus \{$"and"$\} = ||P||$, i. e., each word of the premise is contained in at least one extracted tuple (except "and"),

3. $\forall t \in e(P), ||H|| \not\subseteq ||t||$, i. e., we do not allow situations, where the hypothesis is a shortening of some of the considered tuples,

4. $\forall t, u \in e(P), t \neq u, ||t|| \not\subseteq ||u||$ and $||u|| \not\subseteq ||t||$.

The first condition is obvious. The second one ensures that the information contained in the set of extracted tuples is the same as in the original premise under the natural assumption that a content of a sentence can be fully represented by a set of textual n-tuples for sufficiently high n. By the third condition we want to avoid cases of trivial entailment from one of the multiple premises. The last condition excludes situations when the result $M(P)$ set contains "inclusion sentences" like: `Three men standing on grass` and `Three men standing on grass by the water looking at something on a table.`

The procedure of generating the MPE corpus from an existing NLI one is now straightforward: for each $P - H - L$ triple of NLI corpus check the conditions 3.1 for P, $e(P)$ and H – if satisfied, the obtained $M(P)$-H-L item is added into MPE corpus being created. These items can be further filtered according to different intentions, e. g., we want to deal with items having at least k premises or at most m premises, for instance. Obviously, the quality of the annotation of the original SPE corpus hardly influences the quality of the obtained MPE corpus.

Unlike the MPE corpus from (Lai et al., 2017) where all entailment items contain a set of a *fixed number* of premises (4), we will generally obtain sets of *variable number* of premises.

Reducing the number of trivial inferences by limiting lexical overlap In order to reduce the number of items where the inferences can be done trivially, we may set up a threshold for lexical overlap and consider only such items that the fraction of the number of hypothesis tokens that appear in at least one premise and the number of hypothesis tokens after stopwords removal is lower or equal to a given threshold.

Remark The process can also produce positive (entailment) items where not all premises in the $M(P)$ set take part in the entailment of the hypothesis, i. e., one or more of the premises involved together with the hypothesis form a neutral pair/pairs. This phenomenon will be a subject of further investigations. Nevertheless, in real situations, it is natural to deal with it, as well as with premises such that one extends information provided by another, thus not fulfilling condition 4.

According to (Poliak et al., 2018) again, our proposed corpus can be classified as *automatically recast*, i. e., a corpus that was automatically generated from an existing dataset constructed for a different NLP task and the labeling was done with no or a little manual work. In such cases, some properties of the source corpora may be also transferred into the recasts.

3.2 Quality Evaluation of Our Corpus

NLI corpora are prone to contain several types of *annotation artifacts* (Gururangan et al., 2018). For example, a negation can indicate a contradiction label, mainly in human elicited corpora. Using "generic" words such as "animal" or "instrument" is often typical for the entailment label thanks to common annotators' strategies.

In order to obtain a better insight into the characteristics of the obtained corpus, we are going to investigate the role of occurrence of certain single words in hypotheses when predicting labels and the role of the hypothesis-context relation as well as appearance of different (annotators') patterns.

3.3 Lexical Biases

At first, we are going to focus on words that are strongly indicative of each inference class. In (Gururangan et al., 2018), the authors use pointwise mutual information (PMI) for all words and classes in the training set:

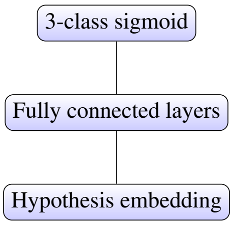
$$PMI(word, class) = \log \frac{p(word, class)}{p(word, .)p(., class)}.$$

The authors use add-100 smoothing to emphasize word-class correlation and select Top 5 words in each class.

In (Poliak et al., 2018), a conditional probability was used:

$$p(l|w) = \frac{count(w, l)}{count(w)}.$$

Figure 1: Overall architecture of the simple hypothesis-only classifier



Then they analogously select Top words for each class (label) l. If $p(l|w)$ is highly skewed across labels, there exists the potential for a predictive bias (Poliak et al., 2018). In this paper, we are going to use this second approach.

3.4 Hypothesis-Only Approach

Annotation artifacts in NLI corpora are common, since annotators, mainly in human elicited corpora development, have different strategies and patterns for generating hypotheses. There are also artifacts that arise from the "hypothesis-context" relation. To model the degree of annotation artifact existence, a classifier that uses only hypotheses for predicting the entailment classes can be trained. In other words, the classifier completely ignoring the information contained in the premise(s) is used.

In (Poliak et al., 2018), the authors call this approach "hypothesis-only" and they used a modified `InferSent` model (Conneau et al., 2017).

In (Gururangan et al., 2018), `fastText` (Joulin et al., 2017), bag-of-words and bigram based model was used, there it is called "premise-oblivious text classifier".

A general architecture of a hypothesis-only classifier is depicted in Figure 1.

Since the outstanding results of BERT model in SPE NLI task[2], we use BERT (Devlin et al., 2018) embeddings.

4 Results and Discussion

The proposed approach for MPE corpus development can be generally used on any single premise NLI annotated corpus in a language where a suitable open IE system is available. To demonstrate it on a concrete dataset, we have chosen the already mentioned SNLI corpus.

4.1 MPE(SNLI) Corpus

As an input for our approach, the training dataset of SNLI was used. From 150.736 *unique* texts/premises, 229.428 n-tuples were extracted, i. e., 1.522 n-tuples per sentence in average. The extraction process was performed by Open IE 5.0[3] (Mausam, 2016). Additional labels provided by the system (like `L:` for location) were removed.

For premise-hypothesis-GoldLabel items of SNLI train, we follow the list of requirements in subsection 3.1. The SNLI corpus contains appox. 2% of items without GoldLabel (marked "-") (Bowman et al., 2015). Even if P-H-L item with L ="-" meets the requirements, it is not added to the corpus being created. Hence the corpus contains only three common labels (entailment, neutral, contradiction) – the number of such items was 96.

Moreover, we set-up a threshold for a lexical overlap according to (Lai et al., 2017) to be equal to 0.5.

The obtained dataset contains 45.622 items. It was then randomly split into train/dev/test datasets containing 32.000, 7.000 and 6.622 items, respectively. Although the SNLI corpus is roughly balanced between the three classes, our corpus contains slightly higher fraction of contradiction labels, mainly because of application of the lexical overlap threshold – high lexical overlap typically indicates the entailment class.

To provide a better idea about the proposed corpus, we provide an example of each label – the premises are syntactically transformed (from t to $s(t)$ in our notation): from textual tuples into sentences, i. e. brackets and semicolons are removed, the first letter is capitalized.

Example 1
Premises:

1. *A white dog with his tongue out is in the snow.*
2. *A brown dog with his tongue out is in the snow.*
3. *A black dog with his tongue out is in the snow.*

Hypothesis: *There are animals outdoors.*
Label: ⇒ Entailment

Example 2
Premises:

[2] See state-of-the-art results on SciTail corpus: https://leaderboard.allenai.org/scitail/submissions/public

[3] https://github.com/dair-iitd/OpenIE-standalone

1. *3 women posing for a picture.*
2. *3 women are sitting down.*

Hypothesis: *The women are smiling.*
Label: ⇒ NEUTRAL

Example 3
Premises:

1. *A baby has food on his face.*
2. *A baby eats.*

Hypothesis: *Baby playing with a dog.*
Label: ⇒ CONTRADICTION

The distribution of labels in the corpus is provided in Table 1.

Table 1: Distribution of labels in MPE(SNLI)

	Train	**Dev**	**Test**
entailment	0.231	0.235	0.231
neutral	0.361	0.362	0.372
contradiction	0.407	0.403	0.396

As already mentioned, we do not require the same number of premises in each corpus item, the distribution of number of premises in the corpus is also provided, see Table 2.

Table 2: Number of premises in MPE(SNLI)

# prem.	2	3	4	5	≥ 6
# items	37765	5013	2294	253	297

4.2 MPE(SNLI) Lexical Biases

In order to select the most characteristic words for each entailment label, we used the conditional probability of a word w w. r. t. the label l. Since there are extremely discriminative words having a very low frequency, we focus only on words that appear at least five times in the training dataset. Table 3, Table 4, and Table 5 present Top 5 words for each class in MPE(SNLI) corpus (training): entailment, neutral, and contradiction respectively together with corresponding values of conditional probability.

These results correspond with our intuition, if we consider the fact that the source of SNLI sentences are mainly photo captions and the fact that the original SNLI corpus was human elicited. Using lexical items that refer to "general" words like *object, similar, multicolored* matches common strategies and patterns when creating entailment pairs. We can also note that the first item in the "contradiction list" is linked to negation. Investigations of annotation strategies and patterns can be a part of the future work.

Table 3: Cond. prob.: l = "entailment"

w	$p(l\|w)$
similar	0.875
facial	0.857
multicolored	0.857
object	0.848
least	0.840

Table 4: Cond. prob.: l = "neutral"

w	$p(l\|w)$
favorite	0.976
tired	0.964
first	0.940
tips	0.9385
tour	0.933

Table 5: Cond. prob.: l = "contradiction"

w	$p(l\|w)$
nobody	0.994
naked	0.971
quietly	0.968
cats	0.938
napping	0.931

4.3 Hypothesis-Only Classifier

Sentence (hypothesis) BERT embeddings were computed using `BERT-as-a-service` application (Xiao, 2018). We have used a pretrained BERT model[4] (12-layer, 768-hidden, 12-heads, 110M parameters). Each hypothesis was encoded as a 768-dimensional vector. The optimal dimension ($d = 100$) of the single hidden layer was obtained by a grid search.

The achieved accuracy reached $\mathbf{0.671}$ on the test dataset, highly above the majority baseline that equals $\mathbf{0.396}$. This result indicates a notable

[4] https://storage.googleapis.com/bert_models/2018_10_18/uncased_L-12_H-768_A-12.zip

presence of annotation artifacts in MPE(SNLI) corpus. Nevertheless, approximately the same accuracy was obtained on the test set of the original SNLI single premise corpus using premise-oblivious `fastText` classifier. We may conclude that proposed corpus achieves a comparable level of annotation artifacts occurrence.

4.4 Conclusion

We have proposed a method of exploiting an open information extraction system for transforming a NLI single premise corpus into a multiple premises entailment (MPE) setting. The method was then applied on SNLI training data and an annotated MPE(SNLI) corpus was obtained. The final corpus will be available at: https://github.com/martinvita/openIE-MPE

4.5 Future Work

This work presents the first steps in creating corpora for MPE task using open IE with a particular application on SNLI data. The further work on the proposed MPE(SNLI) corpus will include extending investigations outlined in this paper, e. g. obtaining deeper insight into lexical bias or investigations about individual entailment classification between hypotheses and individual premises – "elements of M(P)" etc. as a natural continuation of the work.

The keystone of the further work are the investigations of relations among the members of the premise set $M(P)$ together with the development of the entire MPE task . According to our intuition, the MPE should be significantly more difficult than SPE task in the sense that the entailment judgement should be based on a fusion of information contained in the premises. This is also a natural step after the initial work (Lai et al., 2017) – approving the importance of the MPE task.

Both the quality and quantity of extracted tuples – hence also the number of premises and items in the MPE corpus – are strongly influenced by the functionality and quality of the open information extraction system used. Comparing results/outputs of different open IE systems is another prospective field of study.

As we have seen from examples provided in the text, the proposed process led to the MPE corpus with a variable number of premises whereas in the first MPE corpus (Lai et al., 2017), each item contains preciously four premises. In order to prepare a formally compatible corpus, we are interested also in "normalizing" the number of premises using various techniques, e. g. a paraphrase generation for cases when we have less premises than needed, and concatenation in cases when the number of premises exceeds the required number.

Obviously, the proposed process does not rely on certain properties of SNLI, hence it can be straightforwardly applied to other corpora, (e. g. MultiNLI, SciTail etc.), even to NLI single-premise corpora in different languages where open IE tools are available. SNLI is prone to several biases (that are transferred to MPE corpus), thus we can expect the result obtained by applying our procedure on other corpora can lead to more valuable and inspiring results.

A general task when having MPE corpora of suitable quality and volume, is the development of classifiers for MPE task based on different architectures, i. e., general development in MPE field as well as further study on the mutual relationship between the SPE NLI and the MPE NLI task.

Remark Finally, it should be noticed that MPE task is related to NLI with external/background knowledge, which seems to be a promising direction in the field of NLU (Jiang et al., 2018). Having a premise-hypothesis pair and an external/background knowledge that can be formalized in the form of sentences, we can generally add these sentences as additional premises. The key question is obviously the process of selection/recommendation of relevant sentences to become these new additional premises. Clearly, the number of premises needs to be limited. This observation illustrates the importance of the MPE task in the entire NLU field.

References

Asma Ben Abacha and Dina Demner-Fushman. 2016. Recognizing question entailment for medical question answering. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, volume 2016, page 310.

Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *IJCAI*. volume 7, pages 2670–2676.

Luisa Bentivogli, Ido Dagan, and Bernardo Magnini. 2017. The recognizing textual entailment challenges: Datasets and methodologies. In *Handbook of Linguistic Annotation*, Springer, pages 1119–1147.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364* .

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*. Springer, pages 177–190.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324* .

Shan Jiang, Bohan Li, Chunhua Liu, and Dong Yu. 2018. Knowledge augmented inference network for natural language inference. In *China Conference on Knowledge Graph and Semantic Computing*. Springer, pages 129–135.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. pages 427–431.

Daniel Z Korman, Eric Mack, Jacob Jett, and Allen H Renear. 2018. Defining textual entailment. *Journal of the Association for Information Science and Technology* 69(6):763–772.

Alice Lai, Yonatan Bisk, and Julia Hockenmaier. 2017. Natural language inference from multiple premises. *arXiv preprint arXiv:1710.02925* .

Omer Levy and Ido Dagan. 2016. Annotating relation inference in context via question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. volume 2, pages 249–255.

Omer Levy, Torsten Zesch, Ido Dagan, and Iryna Gurevych. 2013. Recognizing partial textual entailment. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. volume 2, pages 451–455.

Mausam Mausam. 2016. Open information extraction systems and downstream applications. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. AAAI Press, pages 4074–4077.

Rodney D Nielsen, Wayne Ward, and James H Martin. 2009. Recognizing entailment in intelligent tutoring systems. *Natural Language Engineering* 15(4):479–501.

Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. 2018. A survey on open information extraction. *arXiv preprint arXiv:1806.05599* .

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*. pages 2641–2649.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. *arXiv preprint arXiv:1805.01042* .

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. pages 885–895.

Martin Víta. 2018. From building corpora for recognizing faceted entailment to recognizing relational entailment. In *Position Papers of the 2018 Federated Conference on Computer Science and Information Systems, FedCSIS 2018, Poznań, Poland, September 9-12, 2018.*. pages 33–38. https://doi.org/10.15439/2018F381.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, pages 1112–1122. http://aclweb.org/anthology/N18-1101.

Han Xiao. 2018. bert-as-service. `https://github.com/hanxiao/bert-as-service`.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2:67–78.