The Multi-genre NLI Corpus 0.9: RepEval Shared Task Preliminary Version Description Paper

Adina Williams
Department of Linguistics
New York University
adinawilliams@nyu.edu

Nikita Nangia
Center for Data Science
New York University
nikitanangia@nyu.edu

Samuel R. Bowman
Department of Linguistics
and Center for Data Science
New York University
bowman@nyu.edu

Abstract

This paper introduces a draft version of the Multi-genre Natural Language Inference (MultiNLI) corpus, a crowd-sourced collection of 433k sentence pairs annotated with textual entailment information. The corpus is modeled on the SNLI corpus, but differs in that it covers a range of genres of spoken and written text, and supports a distinctive cross-genre generalization evaluation.

1 Introduction

This paper describes a preliminary release of the Multi-genre Natural Language Inference (MultiNLI) Corpus, created for the shared task of the the Second Workshop on Evaluating Vector Space Representations for NLP (RepEval 2017). There may be minor updates to this dataset before formal publication and before the conclusion of the shared task evaluation period.

The corpus consists of sentence pairs labeled for the task of *natural language inference* (NLI; Fyodorov et al., 2000; Condoravdi et al., 2003; Bos and Markert, 2005; Dagan et al., 2006; MacCartney and Manning, 2009; Bowman et al., 2015), also known as *recognizing textual entailment* (RTE). The data presented here is separate from, but closely inspired by, the Stanford NLI Corpus (SNLI; Bowman et al., 2015). We follow that work in treating NLI as a balanced three-way classification problem, in the precise definitions of the labels (*entailment*, *neutral*, and *contradiction*), and in our approach to crowd-sourced data collection.

As the name suggests, the corpus consists of sentences derived from ten different sources of text, reflecting ten different genres of written and spoken English. All of the sources are present in test and development sets, but only five are included in the training set.

The shared task has two goals, which correspond to two different ways of using the corpus's genre split. The shared task aims to establish and improve the state of the art for the problem of *sentence encoding*—extracting a real-valued vector to represent the meaning of a sentence. Competitors focusing on this problem should primarily evaluate their systems on the *matched* sections of the development and test sets. These sections contain the same genres of text as the training set.

In addition, the shared task also aims to invite new work on the problem of cross-genre generalization (or *domain transfer*) in sentence level models in NLP. Competitors focusing on this problem should evaluate both on the *matched* and *mismatched* sections of the development and test data. The *mismatched* sections represent genres that do not appear in the training set.

2 Data collection

Our data collection method is modeled closely on that used for SNLI (Bowman et al., 2015), and interested readers should refer to that paper for details. To collect a sentence pair using this method, a crowd worker is presented with a sentence drawn from a source text and prompted to compose a novel sentence which is necessarily true or appropriate in the same situations (yielding a label of *entailment*),

Met my first girlfriend that way.	FACE-TO-FACE **contradiction** C C N C	I didn't meet my first girlfriend until later.
His family had lost a son and a daughter now.	FICTION **entailment** N E E N	The family had children that have passed away.
8 million in relief in the form of emergency housing.	GOVERNMENT **neutral** N N N N	The 8 million dollars for emergency housing was still not enough to solve the problem.
Now, as children tend their gardens, they have a new appreciation of their relationship to the land, their cultural heritage, and their community.	LETTERS **neutral** N N N N	All of the children love working in their gardens.
The Secret Service initiated a number of security enhancements around the White House complex.	NINE ELEVEN **entailment** N E E N	White House gained improvements by enhanced security.
In contrast, suppliers that have continued to innovate and expand their use of the four practices, as well as other activities described in previous chapters, keep outperforming the industry as a whole.	OUP **contradiction** C C C C	The suppliers that continued to innovate in their use of the four practices consistently underperformed in the industry.
I am a lacto-vegetarian.	SLATE **neutral** N N E N	I enjoy eating cheese too much to abstain from dairy.
someone else noticed it and i said well i guess that's true and it was somewhat melodio us in other words it wasn't just you know it was really funny	TELEPHONE **contradiction** C C C C	No one noticed and it wasn't funny at all.
For more than 26 centuries it has witnessed countless declines, falls, and rebirths, and today continues to resist the assaults of brutal modernity in its time-locked, color-rich historical center.	TRAVEL **entailment** E E E E	It has been around for more than 26 centuries.
The father can beget new offspring safe from Macbeth's hand; the son is the palpable threat.	VERBATIM **neutral** N N N N	The son wants to kill him to marry his mom

Table 1: Randomly chosen examples from the development set of our new corpus, shown with their genre labels, and both the selected gold labels and the full set of validation labels (abbreviated) from the individual annotators.

a novel sentence which is necessarily false or inappropriate whenever the first is true (*contradiction*), and one where neither condition applies (*neutral*).

The collection of MultiNLI differs from the collection of SNLI in a few ways:

- We did not recruit workers over Amazon Mechanical Turk, but rather used a prescreened worker pool and presented prompts using the Hybrid platform.
- The wording of the prompts that workers saw varied from genre to genre, rather than being fixed (and specific to the captions genre) in SNLI.
- Premise sentences were collected from the Open American National Corpus distribution augmented with works of fiction (see Appendix) from the public domain, rather than Flickr30k captions in SNLI.
- Only the test and development set examples went through the second *validation* phase of data collection.

3 Data Sources

The majority of the premise sentences presented to crowd workers are drawn from the Open American National Corpus. The Open American National Corpus (OANC) is a permissively-licensed subset of the ANC Second Release. We selected 9 genres out of the 12 provided by the most recent OANC distribution (downloaded: December 2016).

- FACE-TO-FACE: The Charlotte Narrative and Conversation Collection (CNCC) contains 95 narratives, conversations, and interviews documenting life in the Charlotte, North Carolina region in the twentieth century. These communications are part of the Oral History Collections in the Department of Special Collections at J. Murrey Atkins Library, UNC Charlotte.
- TELEPHONE: The OANC selected a portion of University of Pennsylvania's Linguistic Data Consortium Switchboard corpus consisting of 2320 spontaneous conversations (6 minutes long on average), comprising about 3 million words of text (distribution: LDC93S7-T). The speech comes from two-sided, transcribed conversation, spoken by over 500 speakers of both sexes from several major dialects of American English.
- 9/11: This portion consists of The National Commission on Terrorist Attacks Upon the United States' report released on July 22, 2004.
- TRAVEL: Berlitz Travel Guides written by and for Americans were contributed by Langensheidt Publishers. These documents discuss topics relating to vacation travel, including Hotels and Restaurants, History, Where to Go, What to Do, Jungle, and an Introduction.
- LETTERS: The Indiana Center for Intercultural Communication corpus of Philanthropic Fundraising Discourse corpus consists of letters intended to assist in fundraising, including case statements, annual reports grant proposals, and direct mail letters.
- OUP: The Oxford University Press genre contains non-fiction drawn from Oxford University Press publications authored by five American authors (roughly a quarter million words). These works of non-fiction chosen for this project were taken from the earlier portions of the OUP corpus, and are about the textile industry, and child development.
- SLATE: The ANC Slate genre contains 4694 articles from the archives of Slate Magazine, and online publication presenting short articles on popular topics, including new, politics, arts, business, sports, technology, travel, and food. The OANC selection was published between 1996 and 2000, and was written by laypeople.
- VERBATIM: Verbatim is a quarterly about language and linguistics for non-specialist, containing short posts on topics of etymology and language use. The OANC Second Release contains 32 issues of Verbatim from 1990 to 1996.
- GOVERNMENT: The government data in OANC was gleaned from the websites of the

Genre	Train	Dev	Test
SNLI (not incl.)	*550,152*	*10,000*	*10,000*
FICTION	77,348	2,000	2,000
GOVERNMENT	77,350	2,000	2,000
SLATE	77,306	2,000	2,000
TELEPHONE	83,348	2,000	2,000
TRAVEL	77,350	2,000	2,000
9/11	0	2,000	2,000
FACE-TO-FACE	0	2,000	2,000
LETTERS	0	2,000	2,000
OUP	0	2,000	2,000
VERBATIM	0	2,000	2,000
Total (MultiNLI)	**392,702**	**20,000**	**20,000**

Table 2: The number of sentence pairs in each section of the data. The first five genres represent the *matched* section of the development and test sets, and the remaining five represent the *mismatched* section. The test set will be used for the shared task evaluation and is not currently available to the public.

> Environmental Protection Agency, the General Accounting Office, the Japan US Friendship Commission, the Legal Services Corporation, the National Center for Injury Prevention and Control, and the Postal Rate Commission. These data include reports, speeches, letters, and press releases.

We compiled a FICTION genre from various openly available works of contemporary fiction. This genre consists of modern works of short to mid-length fiction (written between 1912 and 2010), spanning multiple genres, including crime, mystery, humor, western, adventure, science fiction, and fantasy. The individual data sources used can be found in the appendix.

Despite the fact that SNLI was collected in largely the same way as MultiNLI, no SNLI examples are included in the distributed MultiNLI corpus. That corpus consists only of sentences derived from image captions from the Flickr30k corpus (Young et al., 2014), and can be treated as a large additional genre (CAPTIONS) dataset.

Note that 353 sentence pairs in the development set (1.8%) have gold labels of '-', indicating a lack of consensus during the validation phase of annotation. These should be ignored in typical model evaluations.

4 Availability, license, and structure

The corpus is freely available for typical machine learning uses, and may be modified and redistributed subject to the terms of the licenses. The data in the FICTION section falls under several licenses, as described below (see Appendix). The remainder of the corpus is released under the Open ANC's license.

The training and development sets are available now at `nyu.edu/projects/bowman/multinli/`. The test set will not be released during the shared task, and test evaluations must be done through the shared task platform.

The corpus can be downloaded in either of two formats, tab separated text and JSON Lines (`jsonl`). Both formats have the following fields for each example:

- `gold_label`: The label to be used for classification. In examples which were rejected during the validation process, the value of this field will be '-'. In typical evaluations, you can ignore these unlabeled examples.
- `sentence1`: The premise sentence for the pair, extracted verbatim from one of the sources described above.
- `sentence2`: The hypothesis sentence for the pair, composed by a crowd worker as a companion sentence for `sentence1`.
- `sentence{1,2}_parse`: Each sentence as parsed by the Stanford PCFG Parser (3.5.2).
- `sentence{1,2}_binary_parse`: The above parse in unlabeled binary-branching format.
- `promptID`: A unique identifier for each premise sentence. Note that a single premise will typically appear in three different sentence pairs, each with a different label.
- `pairID`: A unique identifier for each pair.
- `genre`: The genre of the source text from which `sentence1` was drawn.

- `label[1]`: The label assigned during the creation of the sentence pair. In rare cases for dev and test data, this may be different from `gold_label` if a consensus of annotators during the validation phase chose a different label.

- `label[2...5]`: The four labels assigned to each development and test example during validation. For training examples, these fields will not be filled.

5 Key Statistics

Key size statistics for the distributed corpus can be found in Table 2.

According to the Stanford parser, 90% of premises and 98% of hypotheses are complete sentences, rather than fragments. The mean sentence length is 22 for premise sentences and 11 for hypothesis sentences. The maximum sentence length is 401 for premise sentences and 70 for hypothesis sentences.

6 Baselines and analysis

Simply predicting *entailment* for every example will yield an accuracy of 35.2%/35.4% (mismatched/matched) on the development set and 36.5%/35.6% (mismatched/matched) on the test set.

Additional baseline numbers, including those for human performance and for a selection of techniques shown to work well on SNLI, will be made available during the competition.

In addition, we expect to release an annotated subset of the development data to aid in error analysis.

7 Closing note

The corpus described by this paper is a draft version. We welcome suggestions for ways to better structure or present data of this kind.

Acknowledgments

This work was made possible by a Google Faculty Research Award to Sam Bowman and Angeliki Lazaridou. We also thank George Dahl, the organizers of the RepEval 2016 and RepEval 2017 workshops, and our colleagues at NYU for their help and advice.

References

Johan Bos and Katja Markert. 2005. Recognising textual entailment with logical inference. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP)*. Association for Computational Linguistics, Vancouver, Canada, pages 628–635.

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*. Association for Computational Linguistics, Lisbon, Portugal, pages 632–642.

Cleo Condoravdi, Dick Crouch, Valeria de Paiva, Reinhard Stolle, and Daniel G. Bobrow. 2003. Entailment, intensionality and text understanding. In Graeme Hirst and Sergei Nirenburg, editors, *Proceedings of the HLT-NAACL 2003 Workshop on Text Meaning*. pages 38–45.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine learning challenges. Evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, Springer, New York, NY, pages 177–190.

Yaroslav Fyodorov, Yoad Winter, and Nissim Francez. 2000. A natural logic inference system. In *Proceedings of the 2nd Workshop on Inference in Computational Semantics (ICoS)*.

Bill MacCartney and Christopher D. Manning. 2009. An extended model of natural logic. In *Proceedings of the Eight International Conference on Computational Semantics (IWCS)*. Association for Computational Linguistics, Tilburg, The Netherlands, pages 140–156.

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.

Appendix

The FICTION genre is derived from the following sources:

- *Seven Swords* is an adventure/fantasy novel written by American novelist Mike Shea in 2008 (56,860 words). This work is distributed with the permission of the author, and has a flesch-kincaid reading ease rating of 90.6. The version used for this work was gleaned from manybooks.net. Seven Swords is available under a Creative Commons Share-Alike 3.0 Unported License, and with the explicit permission of the author.

- *Living History* is a historical/science fiction short story comprised of 31,185 words, which was written by Ben Essex (Elliott Gesswell) in 2010. It is about "clones, dinosaurs and the concept of Benjamin Franklin," and has a flesch-kincaid reading ease rating of 80.3. The version used for this work was gleaned from manybooks.net. Living history is available under a Creative Commons License Attribution 3.0 Unported, which allows free sharing and adaptation for all purposes.

- *Password Incorrect* is a humorous 'tech-fiction' short story anthology consisting of 21,646 words, which was written by a Polish-American writer Nick Name (Piotr Kowalczyk) in 2008. It has a has a flesch-kincaid reading ease rating of 62.1. The version used for this work was gleaned from manybooks.net. Password Incorrect is available under a Creative Commons Attribution 3.0 Unported License, which allows free sharing and adaptation for all purposes.

- *Secret Adversary* is the second published mystery/detective fiction novel by British novelist Agatha Christie. It was first published in January 1922 in the United Kingdom. The version used for this work was gleaned from Project Gutenberg. *Secret Adversary* can be freely used in the United States and is not under US copyright. This work may not be freely available for use in other countries.

- *The Mysterious Affair at Styles* is a British mystery/detective fiction novel, which was published in 1912 by Agatha Christie, and follows the exploits of her classic Belgian hero, Hercule Poirot. The version used for this work was gleaned from Project Gutenberg. *The Mysterious Affair at Styles* can be freely used in the United States and is not under US copyright. This work may not be freely available for use in other countries.

- *The Sky is Falling* is a work of short science fiction written in 1954 by American author Lester del Ray. The version used for this work was gleaned from Project Gutenberg, and can be freely used in the United States. The Sky is Falling is no longer covered under US copyright, but may not be available for use in other countries.

- *Youth* is a short work of science fiction written in 1952 by American author Isaac Asimov. The version used for this work was gleaned from Project Gutenberg. *Youth* can be freely used in the United States and is not under US copyright. This work may not be freely available for use in other countries.

- *Rebel Spurs* is a work of American historical/western fiction written in 1962 by Andre Norton. The version used for this work was gleaned from Project Gutenberg. *Rebel Spurs* can be freely used in the United States and is not under US copyright. This work may not be freely available for use in other countries.