FANCY: A Diagnostic Data-Set for NLI Models

Guido Rocchietti, Flavia Achena, Giuseppe Marziano, Sara Salaris, Alessandro Lenci
University of Pisa, Italy
guido.rocchietti@gmail.com, flavia.achena@gmail.com,
marzianogiuseppe@gmail.com, sarasalaris16@gmail.com,
alessandro.lenci@unipi.it

Abstract

We present here FANCY (FActivity, Negation, Common-sense, hYpernimy), a new dataset with 4000 sentence pairs concerning complex linguistic phenomena such as factivity, negation, common-sense knowledge, hypernymy and hyponymy. The analysis is developed on two levels: coarse-grained for the labels of the Natural Language Inference (NLI), that is to say the task of determining whether a hypothesis is true (entailment), false (contradiction), or undetermined (neutral) and fine-grained for the linguistic features of each phenomenon. For our experiments, we analyzed the quality of the sentence embeddings generated from two transformer-based neural models, BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019b), that were fine-tuned on MNLI and were tested on our dataset, using CBOW as a baseline. The results obtained are lower than the performance of the same models on benchmarks like GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) and allow us to understand which linguistic features are the most difficult to understand.

1 Introduction

Nowadays it has become more and more important to understand how much neural models applied to Natural Language Processing can understand about language features.

The **probing task** methodology is a simple but effective approach to address this issue (Conneau et al., 2018). A network is trained on a specific task and then the representations are passed to a classifier. The performance of the classifier is evaluated with a dataset constructed to test the understanding of specific linguistic phenomena. If the classifier performs well, then it can be deduced that the neural embeddings have stored syntactic and semantic knowledge relative to those specific linguisitc phenomena.

One of the most widely used tasks for this approach is Natural Language Inference, in which the model must decide whether a *hypothesis* is an entailment, a contradiction, or simply neutral with respect to the *premise*.

Another approach consists in using benchmarks, i.e. datasets relating to various types of tasks, which are able, on the basis of the results obtained, to provide a general judgment on the performance of the model. Although benchmarks are very useful in evaluating the average performance of models, they are less effective in representing a wide range of linguistic phenomena that the models are able to deal with.

It is in this context that the *challenge sets* are born, (also called *adversarial sets*, *stress sets* or *diagnostic sets*) such as the SNLI (Stanford Natural Language Inference) (Bowman et al., 2015) and the MultiNLI (Multi-genre Natural Language Inference) (Williams et al., 2018). This datasets provide the possibility of more specific evaluation frameworks compared to traditional benchmarks (Belinkov and Glass, 2019): as in the case of the probing task, the aim is to evaluate the quality of linguistic information encoded by vector representations.

For our research we built a diagnostic dataset that addresses key aspects of the human knowledge of lexical and compositional meaning, in order to test the deep semantic abilities of the latest computational models.

In this paper, we introduce FANCY, a dataset with 4,000 different hand-annotated sentence pairs

with inference relation between them. In Section 3 we will briefly present the linguistic phenomena we decided to analyze. In Section 4 we will present the methods of dataset construction and in Section 5 we will discuss the results of the experiments conducted on FANCY.

2 Related Work

Despite the progress made in recent years in the study of vector representations, it is still difficult to understand exactly what kind of linguistic properties they capture. The main approaches used in this area are probing tasks and diagnostic datasets.

A probing task is a classification problem focused on the simple linguistic properties of sentences (Conneau et al., 2018). This approach has been used on a wide variety of linguistic phenomena. The work of Ettinger (2016), for example, focused on semantic role and negation scope: the sentence embeddings used are Skip-Thought (Kiros et al., 2015), Paragram (Le and Mikolov, 2014) and those obtained from the average of GloVe word embeddings (Pennington et al., 2014). Adi et al. (2016) verified whether sentence embeddings are able to encode information such as the order, length and content of words in a sentence. These elements were evaluated on sentence embedding produced by CBOW (Continuous Bag-of-Words) and Encoder-Decoder (ED) models, both pre-trained on Wikipedia.

On the other hand, the importance of challenge sets is demonstrated by the fact that some traditional benchmarks have been equipped, in addition to the standard datasets, with challenge sets dedicated entirely to the NLI task. In fact, both GLUE and SuperGLUE have a diagnostic dataset, consisting of about 1000 pairs of manually constructed sentences involving 30 linguistic phenomena, including anaphora, factivity, negation, redundancy, hyponymy, etc. Similar challenge sets have been developed and described in the publications of Naik et al. (2018), a dataset in which the errors committed related to negation, antinomies and numerical reasoning are also investigated, Glockner et al. (2018), a challenge set created with particular reference to common knowledge and McCoy et al. (2019), an evaluation dataset that contains 30,000 specific examples on which neural models perform incorrect classifications, such as lexical overlap, subsequence, constituent, etc.

3 Linguistic Phenomena

We selected four different kinds of linguistic phenomena to analyze: (1) the *factivity*, which address the truthfulness and the factuality of the events mentioned inside the phrases, (2) the *negation*, which in the English language can be expressed by several terms and situations, (3) *hierarchical relations*, i.e. semantic relations like hypernymy between a general term and a more specific term, and (4) the *common-sense knowledge*, which relates to the shared knowledge among speakers about events and facts concerning the real world.

3.1 Factivity

Factivity is a linguistic phenomenon related to the truthfulness of events or concepts that are mentioned and expressed in a sentence: each event, based on the elements contained in the sentence, can assume a certain degree of certainty.

a. John **thinks** it's raining.

b. John **knows** it's raining.

When a speaker reads the non-factive verb *think* (a.), he understands that the event mentioned in the sentence (*it's raining*) is just a possibility, while he deduces that it's a fact when the factive verb *know* is used (b.).

When we talk of situations and events that occur, have occurred or will surely occur in the world, we present them as facts, while we usually complete our tales using approximations in cases where we do not know whether the things we are talking about have actually happened and we are not completely sure of their certainty. It is in this context that we can observe the phenomenon of *factivity* (Saurí and Pustejovsky, 2012).

3.2 Negation

Negation is a complex phenomenon that characterize human language among all (Horn, 1989). From a logical perspective, it is the opposite of affirmation, which means that the truth value of the statement is reversed by the negative. The main challenge is to identify the *scope* of the negative marker within the sentence, i.e. which element is semantically negated (Jackendoff, 1969). If we consider a sentence such as *Mary does not read carefully*, we can observe that the scope is partial, because the negation refers only to the adverb. Besides the most common *not*, *nobody* and *nothing*, we have taken into account all possible negative cases in the English language.

Negation may be implicit, such as *forget* meaning *not remember*, or affixal in such terms as *illegal* or *dis-agreement*. It could be related to quantifiers, in cases such as *not all veggies are tasty* which contradicts *all veggies are not tasty*. Some sentences can occur with double negative markers, such as *John called neither his father nor his mother*. Moreover, we can observe contrastive negation (McCawley, 1991), in sentences like *John drank not coffee but tea*. So, although characteristic of all languages and frequently used, negation is a complex phenomenon to investigate.

3.3 Hierarchical Relations

In many cases, the entailment relations can occur not only at a sentence level but also at a word level, if we consider the meaning relations that exist between words: these kinds of relations are defined as lexical entailment (Roller, 2017) and they are determined for example by *subtype/type* hierarchical relations such as *hyponymy* (*dog* is hyponym of *animal*) and *troponymy* (*run* is troponym of *move*) (Pustejovsky and Batiukova, 2019). We define the *subtype/type* relation as entailment (*dog* entails *animal*) and the *type/subtype* relation as neutral (*animal* does not entail *dog*) (MacCartney and Manning, 2009). However, the logical relations between lexical elements can be differently projected by the properties (upward monotone, downward monotone and non-monotone) of some semantic functions (*projectivity signatures*) such as restrictive quantifiers (some, any, every, etc.), negation and superlative (MacCartney and Manning, 2014). A function is *upward monotone* if the logical relation between premise and hypothesis is projected without change: the sentence *some parrots talk* entails *some birds talk*. A function is *downward monotone* if it reverses the logical relations between premise and hypothesis: *no fish talk* entails *no carp talk*. A function is *non-monotone* if it projects the logical relation between premise and hypothesis as neutral: *most humans talk* does not entail *most animals talk* (and vice-versa).

3.4 Common-Sense Knowledge

The concept of common-sense is hard to define because it is strictly entangled with the way we humans reason. Even though its definition is controversial, we adopt here what Feldman called *The Standard View* (Feldman, 2003). In his book he defined eleven categories that give us an idea of the things we know as human beings. He stated two different thesis that constitute the Standard View: the first one states that *We know a large variety of things in categories (a)-(k)*[1] and the second one states that *Our primary sources of knowledge are (a)-(f)*[2].

Starting from the types suggested by LoBue and Yates (2011), we grouped common-sense into five macro-categories.

Causal Relations The categories in which the statement of the premise causes the hypothesis statement, e.g. *the man had a bath* entails *the man got wet*: here we can see how the fact that the man took a bath is the cause for him of being wet, hence there is a *Cause/Effect* relation. At the same time the fact that *Mary was married to John* automatically implies *John was married to Mary*, therefore the relation is of *Simultaneous Condition*.

Spatial Relations This category includes sentences that specify the physical position of an agent or an object with respect to someone or something, e.g. the fact that *John is inside his home* contradicts the sentence *John is close to his home* because: in this case, the spatial prepositions *inside* and *close to* cannot subsist at the same time.

Temporal Relations In this category are included texts that specify the time of en event with respect to someone or something, e.g. the fact that *Julius Caesar was assassinated in 44 B.C.* implies that *Julius Caesar died before the birth of Christ*. In this example the reader is supposed to know that B.C. indicates the birth of Christ, which is not trivial.

World Knowledge Relations All the categories that suppose a previous knowledge of the phenomenal or human world, for example all the sentences that suppose a geographic knowledge to be correctly tagged, e.g. *Charles Dickens is buried in Westminster Abbey* implies that *Charles Dickens rests in London* only if we know that Westminster is in London.

Other Relations In this set we put all the categories which are not included in the previous ones (e.g., arithmetic relations and mutually exclusive relations). For example, *On the train, there are 340 passengers and 40 employees* implies that *On the train, there are 380 people* because we know that if there are 340 + 40 people on the train then the total of the people will be 380.

[1] The categories that we know, such as the past, morality, science etc.

[2] He individuated six different sources of knowledge such as perception, memory,reasoning etc.

4 Dataset Construction

The dataset created for the experiments consists in 4000 pairs of sentences that were built manually by the authors, and this is because we decided to only include sentences that were as simple and clear as possible, in order to specifically focus on the linguistic features of the phenomena and to exclude other external factors of complexity that could have affected the performance of the neural models. For the construction of FANCY, we followed the diagnostic dataset schema provided with the SuperGlue[3] benchmark for models evaluation, so all the data were inserted in a tabular framework and tagged with the following columns and labels.

Premise and Hypothesis Are the first two columns of the dataset and indicate which sentence is the premise and which is the hypothesis.

FW and BW These two columns point out which one of the sentences should be used as the premise. For instance, if we find the sentence *Granada is in Spain* as the premise, and *Granada is in Europe* as the hypothesis in the database, the column FW (forward) considers the first as the *premise* and the second as the *hypothesis* while the columns BW (backward) considers the second sentence as the premise and the first as the hypothesis. In both of the columns we inserted the correct output: in the example above, the column FW would contain the tag *entailment*, because the first sentence implies the second one, while the column BW would contain the tag *neutral* because the second sentence does not imply the first one but does not contradict that either.

Phenomenon Category This column is very important for this study because it specifies which kind of feature regarding a particular phenomenon is represented by the sentence pairs.

Phenomenon	E	N	C
Factivity	239	465	296
Negation	410	428	158
Hierarchical	369	475	156
Common-sense	388	254	358

Table 1: Distribution of Entailment (E), Neutral (N) and Contradiction (C) labels.

In Table 1 we can see that FANCY is composed of **1406** pairs of sentences that lead to an entailment, **1622** sets of neutral sentences and **968** contradictions.

[3]https://super.gluebenchmark.com/diagnostics

5 Experiments

In this section, we report the results of the experiments conducted using our dataset FANCY. We tested state-of-the-art models for NLI on the four different linguistic phenomena in the dataset. We selected *bert-base-uncased-MNLI* and *roberta-large-mnli*, both of which were finetuned on the MNLI dataset, and also a baseline model based on CBOW. The BERT and RoBERTa models are based on the Transformer architecture and are available on the Hugging Face web page.[4] For what concerns the CBOW model, it was built using the tensorflow library,[5] with the word embeddings generated by GloVe pretrained with 840 Billions tokens, a vocabulary of 2.2 millions cased words and the resulting word vectors with 300 dimensions.[6] The model was then trained on the MultiNLI dataset, so that all three models were trained on the same data.

Set	BERT	RoBERTa	CBOW
MNLI	84.6	90.2	65.2
Factivity	65.2	74.6	45.1
Negation	70.0	82.0	45.0
Hierarchical	49.7	60.4	37.8
Common-sense	57.0	68.0	41.0

Table 2: Accuracies report.

We tested every model on the examples of FANCY. The results in Table 5 show how the models struggled to address these kind of phenomena, if compared with the results on the MNLI. We can see that the baseline model performed quite poorly on all the subsets of our data. RoBERTa is the best performing one, even though it showed poor performances on linguistic phenomena such as common-sense and hierarchical relations while performing better on factivity and negations.

Label	Error	Tot	%
Possibly Fact	257	416	62
Possibly Counterfact	8	50	16
Fact	27	244	11
Counterfact	32	290	11

Table 3: RoBERTa errors on factivity relations.

In Table 3 we can see the errors that RoBERTa made in labeling examples regarding *factivity*. Most of the errors concern examples where the *hypothesis* gave place to a *Possible fact* and therefore should be tagged as *neutral*.

[4]https://huggingface.co/

[5]https://www.tensorflow.org/

[6]https://nlp.stanford.edu/projects/glove/

Premise	Hypothesis	Gold	Pred.
The man was born in 1950.	The man was 18 in 1968.	E	C
No arrow hit the target.	Not all arrows hit the target.	C	E
Bob believes that Twin Peaks is the best tv show ever.	Twin Peaks is the best tv show ever.	N	E
All seagulls fly.	All birds fly.	N	E

Table 4: Error examples. The column Gold contains the correct tags, while the column Predicted contains the incorrect tags predicted by RoBERTa.

Label	Errors	Tot	%
Negation	116	568	62
Implicit Negation	30	146	16
Contrastive Negation	19	179	10
Partial Negation	16	32	8
Affixal Negation	5	75	3

Table 5: RoBERTa errors on negation relations.

In Table 5 it is evident that the largest number of errors belongs to the *Negation* macro-category. In this case, the sentences contained elements such as quantifiers, modals, temporal adverbs and relative pronouns. Therefore, it appears that the comprehension of negation is more difficult when it is related to these elements.

Label	Errors	Tot	%
Downward Monotone	189	222	48
Upward Monotone	25	138	6
Non-Monotone	62	98	16

Table 6: RoBERTa errors on hierarchical relations.

In Table 6 we can see the errors made by the RoBERTa in dealing with hierarchical relationships. Most errors relate to *Downward Monotone* and *Non-Monotone* sentences.

Label	Errors	Tot	%
Temporal Relation	64	182	19.94
Preconditions	53	146	16.51
World Knowledge	26	60	8.10
Spatial Relation	45	148	14.02
Cause/Effect	24	74	7.48

Table 7: RoBERTa errors on common-sense relations.

In Table 7 we show only the most relevant categories for what concerns the errors committed by the model dealing with *common-sense* and *common-knowledge*.

As we can see, *Temporal Relation*, *Preconditions* and *Spatial Relation* are the most difficult categories for the model to label correctly.

As illustrative examples, in Table 4 are four sentences mislabelled by RoBERTa. We note that the sentences are very simple and easy for human beings to understand.

6 Conclusions

Following a large number of recent studies (Naik et al., 2018), (Glockner et al., 2018), (Belinkov et al., 2019), (Liu et al., 2019a), we also tried to investigate whether the latest neural models were able to understand certain linguistic phenomena. On the one hand, we wanted to test the models on the real understanding of the English language, on the other hand, we wanted to build a fine-grained dataset, which allows a detailed analysis of each phenomenon. We tested two of the the most high-performance models such as BERT and RoBERTa and we observed how they struggle dealing with linguistic features that are quite simple to understand for a human being.

We have shown how the models can better handle phenomena such as *factivity* and *negation* if compared with the results obtained on *hierarchical relation* and *common-sense knowledge*. More in particular, we were able to stress how the state-of-the-art models struggle in dealing with linguistic phenomena that are essential for a correct understanding of the language such as the *possibility* generated by a statement, *temporal relations* between entities, the *negation* when there is a presence of *temporal adverbs* and *relative pronouns* and cases of *downward monotone* sentences. In future developments of our work we could use FANCY in order to perform fine tuning on Transformer-based models with the aim of increasing model performance and inferential capabilities. To do this it would be useful to produce more data, possibly annotated by different people, to test the models developed on different types of natural language. At the same time, the dataset could be implemented with other languages, such as Italian.

References

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Trans. Assoc. Comput. Linguist.*, 7:49–72.

Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Durme, and Alexander Rush. 2019. On adversarial removal of hypothesis-only bias in natural language inference. pages 256–262.

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

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136. Association for Computational Linguistics.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Richard Feldman. 2003. *Epistemology*. Prentice Hall.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics.

Laurence Horn. 1989. *A Natural History of Negation*. University of Chicago Press.

Ray S. Jackendoff. 1969. An interpretive theory of negation. *Foundations of Language*, 5(2):218–241. Publisher: Springer.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors.

Quoc V Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents.

Nelson F. Liu, Roy Schwartz, and Noah A. Smith. 2019a. Inoculation by fine-tuning: A method for analyzing challenge datasets. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692 [cs]*.

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

Bill MacCartney and Christopher D. Manning. 2014. Natural logic and natural language inference. In *Computing Meaning*, volume 47, pages 129–147. Springer Netherlands.

James D. McCawley. 1991. *The Syntactic Phenomena of English*.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

James Pustejovsky and Olga Batiukova. 2019. *The Lexicon*. Cambridge University Press, Cambridge, England.

Stephen Creig Roller. 2017. Identifying lexical relationships and entailments with distributional semantics.

Roser Saurí and James Pustejovsky. 2012. Are you sure that this happened? assessing the factuality degree of events in text. *Computational Linguistics - COLI*, 38:1–39.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in neural information processing systems*, volume 32. Curran Associates, Inc.

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)*, pages 1112–1122. Association for Computational Linguistics.