HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Natural Language Inference – Another Triumph for Deep Learning?

Aarne Talman
PhD Student in Language Technology
aarne.talman@helsinki.fi

November 23, 2017

University of Helsinki
Department of Modern Languages

Outline

Outline

Who am I?

- MSc in Computational Linguistics and Formal Grammar, King's College London, 2007.
- BSc in Philosophy, London School of Economics, 2005.
- 11 years in the IT industry working for Tieto, Nokia, Accenture and Gartner.
- Different roles in software development, project and product management and consulting.
- For the past 5 years working in strategy consulting, advising large high-tech and telecoms companies across Europe on their offering and go-to-market strategies.
- PhD student since 2015, but due to work have not been able to make progress.
- Academic interests: computational semantics, formal grammar, logic and more recently machine learning in natural language processing.
- Live with my wife and 2 daughters (7 and 9) in Munkkivuori, Helsinki.

What am I doing here?

I have been working on a new project plan related to Natural Language Inference. Working title:

“Bridging the Gap Between Deep Learning and Rule-Based Approaches to Natural Language Inference”.

I will (most likely) be employed full time in Jörg’s new project **“Natural Language Understanding using cross-lingual grounding”** from the beginning of March 2018.

Outline

Natural Language Inference

Natural Language Inference (NLI) is the problem of determining whether a natural language hypothesis can be inferred from a natural language premise.

A simple example of such a task would be to determine whether the hypothesis *h* below can be inferred from the premise *p*:

- *p* *So far this week, four mine disasters have claimed the lives of at least 60 workers and left 26 others missing.*
- *h* *Mine accidents cause deaths in China.*

A typical NLI task involves identification of whether such hypothesis-premise pairs are entailments, contradictions or neither.

NLI is relatively easy for humans, but has turned out to be quite hard for computers – even when the data is presented in nicely organised sentence pairs.

Natural Language Inference

NLI has central importance when studying natural language understanding, computational semantics and artificial intelligence more generally.

Cooper et al. argue that inferential ability is a key part of our semantic competence:

> *Inferential ability is not only a central manifestation of semantic competence but is in fact centrally constitutive of it. (Cooper et al., 1996)*

But NLI is not only important from scientific point of view, it also has multiple important applications, e.g.

- question answering, semantic search, automatic summarisation, evaluation of machine translation systems, paraphrasing and sentiment analysis

among many others.

NLI test suites and datasets

There are number of different test suites and datasets developed for NLI. These play a central role in development of NLI approaches as they contain validated premise-hypothesis pairs labeled with information about entailment.

- The **FraCaS** test suite of NLI problems contains 346 NLI examples and is widely used in testing NLI methods (Cooper et al., 1996).
- **Recognizing Textual Entailment** (RTE) is another popular NLI task which was initiated by European Commission's PASCAL project (Dagan et al., 2006).
- A recent addition to NLI datasets is the **Stanford Natural Language Inference** (SNLI) corpus (Bowman et al., 2015a). The SNLI corpus contains 570K sentence pairs labeled with entailment, contradiction and neutral.
- The most recent dataset developed for NLI is the **Multi-Genre Natural Language Inference** (MultiNLI) corpus containing 433k examples from 10 distinct genres (Williams et al., 2017), making it the most comprehensive dataset in terms of variety of styles and topics covered.

Example from MultiNLI

Premise:
And to show just how fast Japan's new rulers were catching on, two punitive expeditions were launched against Korea and China in the grand manner of 19th-century gunboat diplomacy.

Hypothesis:
Japan's new rulers were catching on quickly.

Label:
Entailment

In addition to the above, the dataset contains parse structures for the premises and the hypotheses as well as the associated genre, which for this example was Travel.

Other genres include: Fiction, Government, Letters, 9/11, Telephone, Face-to-Face, etc.

Different NLI Approaches

The current NLI approaches to NLI can be roughly divided into two types:

- **Rule-based approches:** Hypotheses and premises are first translated to some logical formalism and the entailment is evaluated at the level of the logic.
- **Machine Learning approaches:** Machine learning models, e.g. deep neural networks, are trained using NLI datasets and the trained models are used to evaluate inferences.

There has also been some work done on bag-of-words and similar “shallow” approaches.

Rule-based approaches

There are multiple different rule-based approaches which vary based on the logical formalism they use.

They all provide a deep semantic analysis for the statements and use logical systems to validate entailment relations.

Some notable approaches in this tradition include:

- **MacCartney (2009):** Natural Logic
- **Bos and Markert (2006):** Combinatory Categorial Grammar (CCG)
- **Bernardy and Chatzikyriakidis (2017):** Modern Type Theory

Natural Logic: 7 basic semantic relations

Image source: Bill MacCartney
https://nlp.stanford.edu/~wcmac/papers/20150410-UPenn-NatLog.pdf

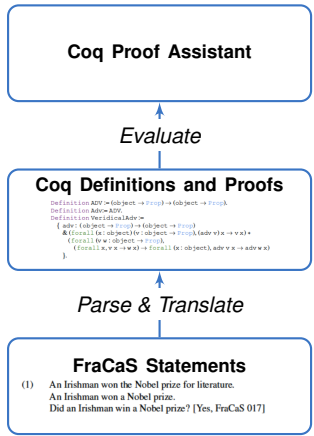
Bernardy and Chatzikyriakidis (2017) implementation of Modern Type Theoretic approach using Coq Proof Assistant

Based on Grammatical Framework and Coq (an interactive proof assistant).

As Coq is an implementation of Modern Type Theory it provides a convenient tools for implementing natural language inferences.

Manually translated parsed FraCaS examples into Coq statements and evaluated 5 sections of FraCaS (174 examples) by using the Coq Proof Assistant.

Were able to improve the state of the art (for FraCaS) by 14 percentage points, obtaining 83% accuracy.



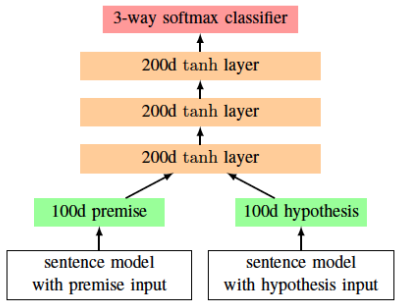
Deep learning approaches

Recently, since the publication of the SNLI corpus, there has been increased interest in various deep learning approaches to NLI.

The large size of the SNLI corpus makes it now feasible to develop deep neural network models for NLI.

In most of the deep learning models used for NLI, both the hypothesis and the premise are represented as a sentence embedding vector. The vectors are then used as an input for a multi-layer neural network (multi-layer perceptron).

A neural network classification architecture (Bowman et al., 2015a):



Deep learning approaches

Some of the neural network models that have been used with SNLI:

- Bowman et al. (2015a) report an accuracy of 77.6 when using long short-term memory (LSTM) networks to achieve the sentence embeddings.
- Rocktäschel et al. (2015) achieve an accuracy of 83.5 when using LSTM word-by-word attention model.
- Wang and Jiang (2015) achieve an accuracy of 86.1 when using a LSTM with word-by-word matching of hypothesis with the premise (*m*LSTM).

A simplified comparison of the rule-based and deep learning approaches

Rule-Based Approaches

+ Provides explanation for the semantic relations involved in entailment and contradiction.
+ Can handle almost any type of semantic relation, if programmed to do so (depends on the chosen formalism).

- Highly fragile and prone to errors.
- Not able to generalise to new types of sentences.
- Manual translation to a logical formalism required.
- Dependent on the test suites.

Deep Learning Approaches

+ More robust than rule based approaches.
+ Can easily handle huge datasets.
+ Generalise better to new sentences (without explicit programming).

- Doesn't provide explanation for the semantic relations involved.
- Dependent on the test suites.

Christopher D. Manning: Computational Linguistics and Deep Learning

[W]hy NLP need not worry about deep learning: [. . .] Our field is the domain science of language technology; it's not about the best method of machine learning – the central issue remains the domain problems. The domain problems will not go away. (Manning, 2015, p. 702)

I would encourage everyone to think about problems, architectures, cognitive science, and the details of human language, how it is learned, processed, and how it changes, rather than just chasing state-of-the-art numbers on a benchmark task. (Manning, 2015, p. 706)

Outline

What would a good NLI system look like?

Some key characteristics of scientifically motivated NLI system would include:

1. Highly **accurate** in identifying entailment, contradiction and neutral

What would a good NLI system look like?

Some key characteristics of scientifically motivated NLI system would include:

1. Highly **accurate** in identifying entailment, contradiction and neutral
2. **Generalise** well to new cases.

What would a good NLI system look like?

Some key characteristics of scientifically motivated NLI system would include:

1. Highly **accurate** in identifying entailment, contradiction and neutral
2. **Generalise** well to new cases.
3. **Robust** (i.e. not fragile and easily prone to errors)

What would a good NLI system look like?

Some key characteristics of scientifically motivated NLI system would include:

1. Highly **accurate** in identifying entailment, contradiction and neutral
2. **Generalise** well to new cases.
3. **Robust** (i.e. not fragile and easily prone to errors)
4. Provides **explanation to the domain problems** of linguistics related to natural language inference (e.g. why does the system identify something as entailment)

What would a good NLI system look like?

Some key characteristics of scientifically motivated NLI system would include:

1. Highly **accurate** in identifying entailment, contradiction and neutral
2. **Generalise** well to new cases.
3. **Robust** (i.e. not fragile and easily prone to errors)
4. Provides **explanation to the domain problems** of linguistics related to natural language inference (e.g. why does the system identify something as entailment)
5. Possibility to be **applied to any textual input** (i.e. identifies and evaluates valid inferences and contradictions from any text)

What would a good NLI system look like?

Do the current deep learning approaches have these capabilities?

1. Highly **accurate** in identifying entailment, contradiction and neutral. (yes)
2. **Generalise** well to new cases. (yes)
3. **Robust** (i.e. not fragile and easily prone to errors). (yes?)
4. Provides **explanation to the domain problems** of linguistics related to natural language inference (e.g. why does the system identify something as entailment). (no)
5. Possibility to be **applied to any textual input** (i.e. identifies valid inferences and contradictions from any text). (no)

NLI – Another Triumph for Deep Learning?

1. Highly **accurate** in identifying entailment, contradiction and neutral
2. **Generalise** well to new cases.
3. **Robust** (i.e. not fragile and easily prone to errors)
4. Provides **explanation to the domain problems** of linguistics related to natural language inference (e.g. why does the system identify something as entailment)
5. Possibility to be **applied to any textual input** (i.e. identifies valid inferences and contradictions from any text)

Without these capabilities we cannot claim that deep learning has won against the alternative approaches

Outline

Potential directions for future research...

1. Highly **accurate** in identifying entailment, contradiction and neutral
2. **Generalise** well to new cases.
3. **Robust** (i.e. not fragile and easily prone to errors)
4. Provides **explanation to the domain problems** of linguistics related to natural language inference (e.g. why does the system identify something as entailment)
5. Possibility to be **applied to any textual input** (i.e. identifies valid inferences and contradictions from any text)

The starting point for my research...

My goal: To develop a more linguistically motivated approach to NLI without losing the benefits of the deep learning approaches.

Potential directions for future research...

1. Combining rule-based and deep learning approaches to NLI
2. Utilitising cross-lingual grounding to enhance deep learning approaches to NLI
3. ...

1. Combining rule-based and deep learning approaches

There has been limited attempts to combine rule-based appraoches and deep learning in the NLI setting.

Bowman et al. (2015b) studied two recursive neural network models: tree-structured neural networks (Tree-RNNs) and tree-structured neural tensor networks (Tree-RNTNs) and their ability to learn to identify logical relationships in Natural Logic formalism of MacCartney and Manning (2009).

Bowman et al.'s approach has been recently replicated by Veldhoen and Zuidema (2017). They found that the system was able to learn some logical relations but failed to learn others, e.g. relationships between quantifiers.

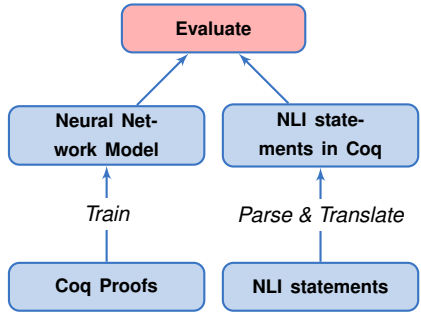
1. Combining rule-based and deep learning approaches

I am working with Stergios Chatzikyriakidis (CLASP, University of Gothenburg) to build a system for learning to identify valid Coq proofs.

The plan is to train a deep neural network using publicly available validated Coq proofs as training data.

The model would be tested initially on the FraCaS examples that have already been translated to Coq by Bernardy and Chatzikyriakidis (2017).

Conceptual illustration of the planned approach



Limitation: Still relies on manual translation to the logical formalism (but maybe there's a way to automate that?)

2. Utilitising cross-lingual grounding to enhance deep learning approaches

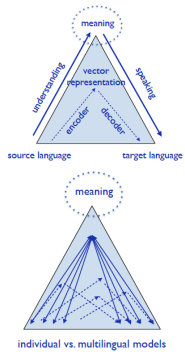
Will work with Jörg Tiedemann on his new project: *Natural Language Understanding using cross-lingual groundin*.

The goal is to build language-independent abstract meaning representations by training neural networks with massively parallel multilingual datasets.

The hypothesis is that by increasing the linguistic diversity in these models, it is possible to resolve language-internal ambiguities and ultimately get closer to a “universal meaning representation”.

These models should help with Neural Manchine Translation tasks, but **can they also help with Natural Language Inference?**

Conceptual illustrations



Source: Jörg Tiedemann

2. Utilitising cross-lingual grounding to enhance deep learning approaches

Initial complication:

- Using massively parallel corpora of human translations might give you **meaning representations of individual sentences**.
- Learning whether natural language inferences are valid involves learning the **connections between the premises and the hypotheses** involved in the NLI task.

2. Utilitising cross-lingual grounding to enhance deep learning approaches

There might still be a way to do this. The plan is to start from the basics.

- **Equivalence:** Testing if two sentences are equivalent shouldn't differ too much from testing if a translation is accurate.

Consider this very much simplified example:

English - Finnish:

- *A cat sleeps on a sofa.*
- *Kissa nukku sohvalla.*

English - English:

- *A cat sleeps on a sofa.*
- *A cat sleeps on a couch.*

2. Utilitising cross-lingual grounding to enhance deep learning approaches

What about other entailment relations?

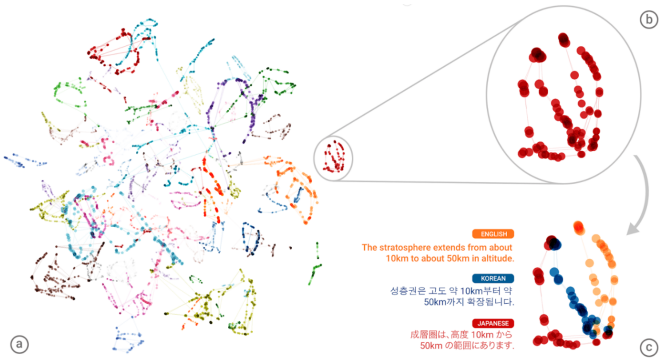
Intuitively these could be assessed in terms of **“similarity”**:

- The meaning representations of two sentences, where one is entailed by the other, should be (somehow) “more similar” to each other as compared to two completely unrelated sentences.

But how to measure similarity?

2. Utilitising cross-lingual grounding to enhance deep learning approaches

Johnson et al. (2016) compared the activations of the network during translation for different sentences. . .



(a) The points with the same colour represent sentences with the same meaning.

(b) Sentences with the same meaning seem to form clusters.

(c) Different colours in this cluster represent sentences in different languages.

Johnson et al. (2016), *Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation*

Thank You!

Questions & Feedback?

References I

Bernardy, J.-P. and Chatzikyriakidis, S. (2017). A type-theoretical system for the fracas test suite: Grammatical framework meets coq. Submitted.

Bos, J. and Markert, K. (2006). Recognising textual entailment with robust logical inference. In Quiñonero-Candela, J., Dagan, I., Magnini, B., and d'Alché Buc, F., editors, *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 404–426, Berlin, Heidelberg. Springer Berlin Heidelberg.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015a). 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.

References II

Bowman, S. R., Potts, C., and Manning, C. D. (2015b). Recursive neural networks can learn logical semantics. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, Stroudsburg, PA. Association for Computational Linguistics.

Cooper, R., Crouch, D., van Eijck, J., Fox, C., van Genabith, J., Jan, J., Kamp, H., Milward, D., Pinkal, M., Poesio, M., Pulman, S., Briscoe, T., Maier, H., and Konrad, K. (1996). Using the framework. Deliverable D16, FraCaS Project.

Dagan, I., Glickman, O., and Magnini, B. (2006). The pascal recognising textual entailment challenge. In Quiñonero-Candela, J., Dagan, I., Magnini, B., and d'Alché Buc, F., editors, *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

References III

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F. B., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2016). Google's multilingual neural machine translation system: Enabling zero-shot translation. *CoRR*, abs/1611.04558.

MacCartney, B. (2009). *Natural Language Inference*. PhD thesis, Stanford, CA, USA. AAI3364139.

MacCartney, B. and Manning, C. D. (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.

Manning, C. D. (2015). Computational linguistics and deep learning. *Comput. Linguist.*, 41(4):701–707.

Rocktäschel, T., Grefenstette, E., Hermann, K. M., Kočiský, T., and Blunsom, P. (2015). Reasoning about entailment with neural attention. In *arXiv preprint arXiv:1509.06664*.

References IV

Veldhoen, S. and Zuidema, W. (2017). Can neural networks learn logical reasoning? In *LAML*. Forthcoming.

Wang, S. and Jiang, J. (2015). Learning natural language inference with LSTM. *CoRR*, abs/1512.08849.

Williams, A., Nangia, N., and Bowman, S. R. (2017). A broad-coverage challenge corpus for sentence understanding through inference. arXiv preprint 1704.05426.