

Understanding and Improving Neural Models for Natural Language Inference using External Resources

Verständnis und Verbesserung Neuronaler Modelle für Natural language inference
Master-Thesis von Max Glockner
Mai 2018



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Fachbereich Physik
Institut für
Festkernphysik
Speerspitze der Elite

Angewandte

Understanding and Improving Neural Models for Natural Language Interference using External Resources

Verständnis und Verbesserung Neuronaler Modelle für Natural language inference

Vorgelegte Master-Thesis von Max Glockner

1. Gutachten: Prof. Dr. Iryna Gurevych
2. Gutachten: Dr. Yoav Goldberg
3. Gutachten: Dr. Andreas Rücké

Tag der Einreichung:

Bitte zitieren Sie dieses Dokument als:

URN: urn:nbn:de:tuda-tuprints-12345

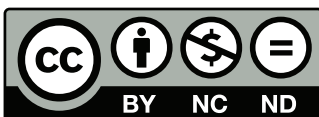
URL: <http://tuprints.ulb.tu-darmstadt.de/1234>

Dieses Dokument wird bereitgestellt von tuprints,

E-Publishing-Service der TU Darmstadt

<http://tuprints.ulb.tu-darmstadt.de>

tuprints@ulb.tu-darmstadt.de



Die Veröffentlichung steht unter folgender Creative Commons Lizenz:

Namensnennung – Keine kommerzielle Nutzung – Keine Bearbeitung 2.0 Deutschland

<http://creativecommons.org/licenses/by-nc-nd/2.0/de/>

Erklärung zur Abschlussarbeit gemäß §23 Abs. 7 APB der TU Darmstadt

Hiermit versichere ich, J. Walker, die vorliegende Master-Thesis ohne Hilfe Dritter und nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die Quellen entnommen wurden, sind als solche kenntlich gemacht worden. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Mir ist bekannt, dass im Fall eines Plagiats (§38 Abs. 2 APB) ein Täuschungsversuch vorliegt, der dazu führt, dass die Arbeit mit 5,0 bewertet und damit ein Prüfungsversuch verbraucht wird. Abschlussarbeiten dürfen nur einmal wiederholt werden.

Bei der abgegebenen Thesis stimmen die schriftliche und die zur Archivierung eingereichte elektronische Fassung überein.

Bei einer Thesis des Fachbereichs Architektur entspricht die eingereichte elektronische Fassung dem vorgestellten Modell und den vorgelegten Plänen.

Datum / Date:

Unterschrift / Signature:

Abstract

TODO in English...

Zusammenfassung

TODO in Deutsch...

List of abbreviations

biLSTM bidirectional Long-Short-Term Memory Network

BoW Bag of Words

ESIM Enhanced Sequential Inference Model

HIT Human Intelligence Task

IE Information Extraction

IR Information Retrieval

KIM Knowledge-based Inference Model

LSTM Long-Short-Term-Memory

MLP Multi Layer Perceptron

MultiNLI MultiGenre Natural Language Inference Corpus

NLI Natural Language Inference

NLP Natural Language Processing

NLU Natural Language Understanding

POS Part of Speech

OANC Open American National Corpus

QA Question Answering

RNN Recurrent Neural Network

RTE Recognizing Textual Entailment

SD Standard Deviation

SICK Sentences Involving Compositional Knowledge

SNLI The Stanford Natural Language Inference Corpus

WSD Word Sense Disambiguation

YAGO Yet Another Great Ontology

Contents

1	Introduction	5
2	Theoretical Background	6
2.1	Natural Language Inference	6
2.2	Lexical Semantic Relations	6
2.2.1	Synonymy and antonymy	6
2.2.2	Hypernymy	7
2.2.3	Holonymy	7
2.3	Shortcut-Stacked-Encoder and Residual Encoder	7
2.3.1	Sentence Encoding for Shortcut-Stacked-Encoder	8
2.3.2	Classification	9
2.3.3	Training	9
2.3.4	Residual Encoder and Reimplementation Variants	9
3	Related Work	11
3.1	External Resources	11
3.1.1	WordNet	11
3.1.2	Wikipedia	12
3.1.3	Derived from multiple Knowledge Bases	12
3.2	Datasets for NLI	12
3.2.1	SNLI	12
3.2.2	MultiNLI	14
3.2.3	SciTail	15
3.3	Neural Models for NLI	16
3.3.1	Sentence Encoding Models	16
3.3.2	Inter-Sentence-Attention-based Models	17
3.4	Integration of external Resources into Neural Networks	17
3.4.1	Improving word-embeddings	17
4	Understanding Shortcut-Stacked-Encoder	19
4.1	Motivation	19
4.2	Insights on the sentence representation	19
4.2.1	Approach	19
4.2.2	Detection of relevant dimensions	20
4.2.3	Female and male dimensions	24
4.2.4	Other semantic dimensions	28
4.2.5	Syntactic dimensions	28
4.3	Insights on the sentence alignment	30
4.3.1	Alignment analysis on a single sample	30
4.3.2	Approach for a general alignment understanding	33
4.3.3	Entailment analysis	34
4.3.4	Neutral and contradiction analysis	35
4.3.5	Summarizing the insights onnMax-pooled sentence representations	36
4.4	Identification of missing knowledge	36
4.4.1	Approach	37
4.4.2	Quantitative results	37
4.4.3	Conclusions	38
5	Additional SNLI test-set	39
5.1	Goal of the new test-set	39
5.2	Dataset	40
5.2.1	Creation of adversarial samples	40
5.2.2	Validation	42
5.3	Evaluation	44
5.3.1	Experimental setup	44
5.3.2	Models with external knowledge	45

5.3.3	Results	45
5.4	Analysis	46
5.4.1	Accuracy by category	46
5.4.2	Impact on the word embeddings	47
6	Approaches to incorporate WordNet information	49
6.1	Extraction of WordNet data	49
6.2	Integrating information into word-embeddings	49
6.2.1	Motivation	49
6.2.2	Concatenating pre-trained word-embeddings	49
6.2.3	Concatenation categorical information	49
6.2.4	Analysis	49
6.3	Multitask Learning	49
6.3.1	Motivation	49
6.3.2	Architecture	49
6.3.3	Approaches	49
6.3.4	Analysis	49
6.3.5	Evaluation	49
7	Conclusion	50

1 Introduction

In recent years neural networks again gained a lot of popularity for many machine learning tasks, including the field of Natural Language Processing (NLP). While previous generation solutions heavily depended on handcrafted features, these models are capable of learning meaningful feature representations automatically (Bengio et al., 2013), thus avoiding time consuming process of feature-engineering. For the most part, neural networks solely rely on distributed word representations, also known referred to as word-embeddings, like word2vec (Mikolov et al., 2013a) or GloVe (Pennington et al., 2014) and typically learn fixed-length dense vector representations for the input text. While they provide strong generalization capabilities, they fail to capture simple world-knowledge (Celikyilmaz et al., 2010) and even have trouble differentiating between mutually exclusive words, if they generally are used in similar contexts (Vulić et al., 2017). As opposed to that, traditional approaches extensively made usage of lexical resources containing relational and factual information about words and entities, thus providing a huge amount of ready-to-use knowledge bases. Intuitively, combining both worlds by integrating existing knowledge bases into neural networks should even further improve these models, due to a more sophisticated Natural Language Understanding (NLU). This is analogous to the way humans understand text, by having a solid understanding of the world, that influences the subjective interpretation of every word within a sentence. Given the sentence “The official language in the USA is English.” an average human can conclude that the official language of *New York* also is English, knowing that *New York* is within the *USA*.

To improve the NLU of neural models for NLP we address the mentioned problems by analysing the sentence representations of a state-of-the-art model and identifying knowledge that is captured or not captured using state-of-the-art strategies without external resources. We show those state-of-the-art models are limited in their generalization ability and fail to capture simple inferences. To overcome this problem, we evaluate how additional knowledge from external resources could be inferred neural networks. While our aim is to provide generally applicable results, we base our experiments on the task of Natural Language Inference (NLI) (Bowman et al., 2015), also known as Recognizing Textual Entailment (RTE) (Dagan et al., 2006). As this is known to be a fundamental task for NLU (MacCartney and Manning, 2007), insights gained here can improve other tasks of NLP that indirectly depend on it.

Structure of the Thesis

While we explain relevant techniques and concepts, we expect the reader to have a basic understanding of common machine-learning practices, neural networks, including basic network architectures like Long-Short-Term-Memory (LSTM) or Recurrent Neural Network (RNN), and NLP in general. This thesis is structured in the following manner:

- Section §2 is used to give definitions for NLI and relevant word-relations. We further give a detailed description about the architecture and training of the state-of-the-art model, that we use throughout all our experiments.
- In section §3 we introduce recently published relevant datasets for NLI and discuss several strategies proved to be successful. In addition we show a selection of lexical resources that contain relevant information to improve the NLU of neural models and various strategies that have been applied to integrate them.
- We analyse how the information of a natural language text is encoded within the sentence representation of a neural model and give insights on how the model uses it in section §4.
- We derive a new testset from a major dataset for NLI, demonstrating the poor generalization abilities of state-of-the-art models in section §5.
- Based on the new data we evaluate whether external resources are helpful for the task using advanced embeddings and multitask-learning.

2 Theoretical Background

This section gives an overview of the task that is used within this work (§2.1) and relevant datasets regarding this task (§3.2).

2.1 Natural Language Inference

NLI (Bowman et al., 2015) deals with the problem to identify, whether one piece of natural text, namely the *hypothesis*, can be inferred from another piece of text, namely the *premise*. The hypothesis h is said to be entailed by the premise p if a human reader would conclude that the hypothesis is true, given the fact that the premise is true. Therefore it differs from strict logical inference. While in NLI a high plausability for the premise to imply the hypothesis, based on the human judgement, is sufficient, the latter one strives to achieve certainty (Dagan et al., 2009). NLI essentially breaks down to an alignment problem (MacCartney et al., 2008). Given the sentence pair

Premise: *Donald Trump is eating his cheeseburger in his bedroom.*
Hypothesis: *The president of the United States is snacking a cheeseburger in the White House.*

the model is required to correctly align *Donald Trump* with *The president of the United States*, *eating* with *snacking* and have information that his *bedroom* is within the *White House*. Here it can be seen, how the system would not only need to cope with different ways of expressing the same meaning, due to the nature of language, but also is required to access and process factual information, that is commonly known to an average human.

Relatedness to other NLP tasks

While NLI clearly is central to reasoning capabilities, it is very fundamental and applicable to a large variety of NLP as the ability to recognize textual entailment is a fundamental and necessary problem towards real NLU (MacCartney and Manning, 2007; Bos and Markert, 2005). Many NLP applications such as Question Answering (QA), Summarization Information Extraction (IE) implicitly depend on this ability, as the huge variability of possible expressions for the same meaning it is a core phenomenon of natural language (Dagan et al., 2009). All three tasks require the model to infer that the target meaning of interest can be inferred from any other variant of textual expression. For QA it is the identification of a correct answer, for summarization, the complete summary needs to be implied by the original text and similarly, while redundant sentences expressing the same meaning should be omitted. Similarly IE, especially if using multiple documents, needs to infer, whether two variants of text contain the same information. Even simple paraphrasing can be broken down to a lexical inference problem with mutual entailment between p and h . As end applications for NLP in addition to NLU need to solve another complicated machine-learning task it is hard to compare and directly improve their NLU capabilities. Thus, one of the main purposes of NLI, being a very basic problem towards NLU, serves as a benchmark with any improvements helping a large variety of high level tasks (Williams et al., 2017; Cooper et al., 1996; Bos and Markert, 2005; Dagan et al., 2006).

2.2 Lexical Semantic Relations

max: **TODO:** lexical inference (Shwartz et al., 2015) and being the semantic relation between two terms (bi/unidirectional) Lexical relations describe the relationship between words¹ whereas *Lexical Semantic Relations* specifically indicate relations referring to the meaning of the word. (Murphy, 2003) and have shown to be helpful for detecting lexical inferences (Dagan et al., 2009). We define the following relations based on those of Jurafsky and Martin (2008). One key characteristic of natural language is ambiguity, that is also present in lexical semantics as words may have several meanings or *senses*². To deal with this phenomenon, lexical semantic relations are defined between senses rather than words. For the sake of simplicity for the most part we follow a naive approach in the following chapters of assuming the most dominant sense of a word, when referring to it. Specifically we define *Synonymy*, *Antonymy*, *Hypernymy* and *Holonymy*, the latter two relations are visualized³ in Figure 1.

2.2.1 Synonymy and antonymy

Synonymy is a symmetric relationship between two senses or two words. Two senses of two different words are said to be synonyms, if they have the same or nearly the same meaning. Synonymy between words holds, if one word can

¹ While there is no single definition for *word*, we refer assume a single word to have the same surface form and the same lemma.

² The phenomenon of words having multiple senses is called *homonymy*, if both senses show indicate no relation but share the surface form like “bank” (financial institution) and “bank” (sloping mound). If those senses are semantically related like “milk” (take milk from female mammals) and “milk” (like cow’s milk), the relationship is called *polysemy* (Jurafsky and Martin, 2008)

³ This is for illustration purposes only and we only added some relevant relations between the entities, more relation are possible. For instance, the holonym relationship would of course hold between *head* and any other *animal*.

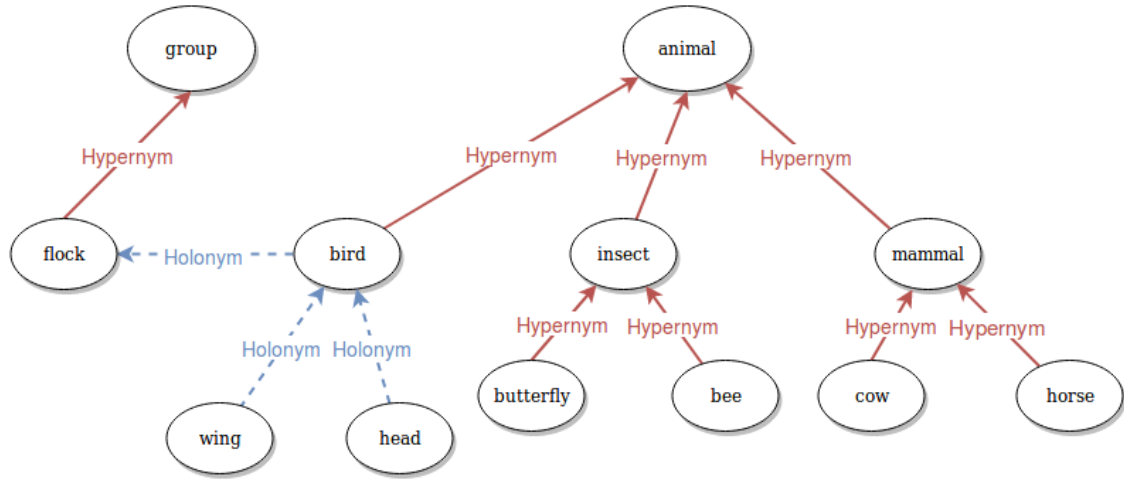


Figure 1: A sample ontology of animals to illustrate the lexical relations *Hypernymy* and *Holonymy*.

be replaced by the other word in any sentence without changing the meaning of the sentence. True synonyms are rare, as most words at least have subtle differences in their meaning or are used within different contexts. We thus follow common practice and loosen the strict definition by referring to synonymys if they have approximately similar meanings. Like synonymy, antonymy is a symmetric relationship between senses, however having the opposite meaning, which might be caused by a binary opposition like “opened/closed”, by different ends on some scale like “hot/cold” or by directional change like “upwards/downwards”. Since antonyms semantically are identical in all other aspects with synonyms, these relations are hard to distinguish from each other automatically.

2.2.2 Hypernymy

Hypernymy (or Hyponymy) is an asymmetric relation between two senses and also referred to as the **is-a** relation. The more specific sense (e.g. *bee*) is called a hyponym of the more general sense (e.g. *insect*), which is called hypernym. Jurafsky and Martin (2008) give a formal definition for Hyponymy in terms of entailment:

“[...] a sense *A* is a hyponym of a sense *B* if everything that is *A* is also *B* and hence being an *A* entails being a *B*, or $\forall x A(x) \Rightarrow B(x)$.” (Jurafsky and Martin, 2008)

Hypernymy is in most cases transitive, thus if a *cow* is hyponym of *mammal* and *mammal* is a hyponym of *animal*, *cow* is also a hyponym of *animal*. an important phenomenon for this thesis are two words, sharing a close hypernym. In Figure 1, *bee* and *butterfly* share the hypernym *insect*, we refer to them as co-hyponyms.

2.2.3 Holonymy

Holonymy or Meronymy refers to the **part-whole** relation. In the illustration of Figure 1, the *wing* is a part of a *bird* and a *bird* is a part of a *flock*. We say that a *bird* is a meronym of *flock*, while *flock* is the holonym of *bird*. As opposed to Hypernymy, this asymmetric relation is not automatically transitive. While a *flock*⁴ obviously consists of several birds, in this case *birds* is generally not replaceable with *heads*.

2.3 Shortcut-Stacked-Encoder and Residual Encoder

We conducted most of our experiments with the Shortcut-Stacked-Encoder (Nie and Bansal, 2017) and the recently adapted version to the Residual Encoder for NLI. They achieve state-of-the-art results⁵ for two large datasets⁶ for NLI and follow the Siamese Architecture, originally introduced by Bromley et al. (1994), thus first encoding *p* and *h* using the same sentence encoder with shared weights into fixed length sentence representations and then predicting the entailment label from the combination of both representations using an additional Multi Layer Perceptron (MLP).

⁴ This is an example for polysemy, as *flock* may refer to a group of birds, but also to a group of e.g. sheep. In this case we assume the sense of a group of birds.

⁵ Considering models for SNLI without inter-sentence attention.

⁶ These are explained in detail in §3.2

2.3.1 Sentence Encoding for Shortcut-Stacked-Encoder

The key novelty of this approach is the way sentence representations are created using a three-layer bidirectional Long-Short-Term Memory Network (biLSTM) with shortcut connections and row-wise max-pooling. An overview of this architecture is given in Figure 2. Due to the arbitrary length of words in textual input a widely used strategy to encode

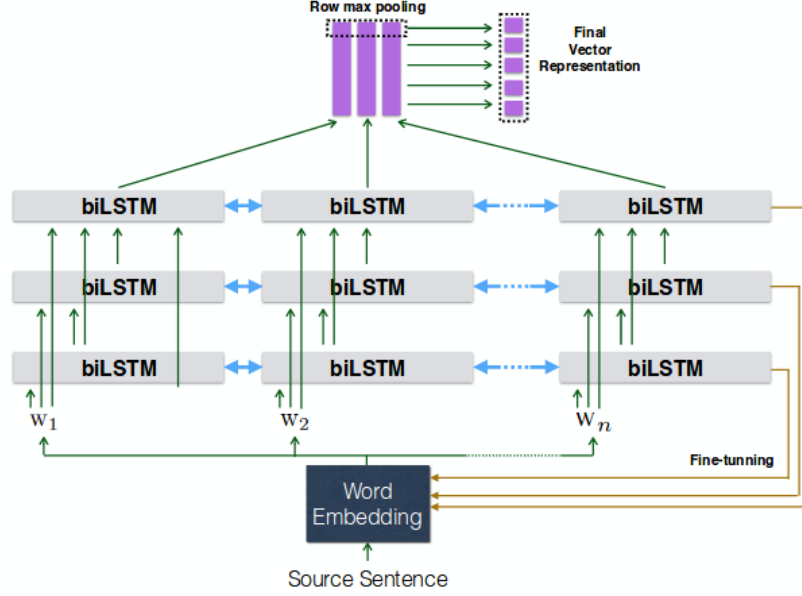


Figure 2: The architecture of the sentence-encoding component within the Shortcut-Stacked-Encoder, taken from Nie and Bansal (2017).

variable length inputs to fixed length vectors using LSTM (Hochreiter and Schmidhuber, 1997) or the bidirectional variant biLSTM (Graves and Schmidhuber, 2005). Essentially these components learn with the use of gates what information to keep and forget at a given point in time. By sequentially going through a sentence in one or two directions respectively, are capable of exploiting word-order and take context into account.

The main difference of the Shortcut-Stacked Sentence-Encoder to typical approaches of a multi-layer biLSTM model is that the input to the biLSTM in a following layer is not only the output of the previous layer, but the output of *all* previous layers, together with the word embeddings. In the first step, the embedding layer maps each word ω_i of the source sentence ($\omega_1, \omega_2, \dots, \omega_n$) to a d -dimensional word vector $w_i \in \mathbb{R}^d$. According to Nie and Bansal (2017) we denote x_t^i to be the input of the i th biLSTM at timestep t . Naturally the input to the first layer are the word-embeddings itself, thus:

$$x_t^1 = w_t \quad (1)$$

In all biLSTM with $i > 1$ the input is the concatenation of all intermediate inputs of previous layers at the timestep t together with the initial word embeddings. Let $[\]$ denote the vector concatenation and h_t^i be the output of the i th biLSTM at timestep t , this leads to:

$$x_t^i = [w_t, h_t^{i-1}, h_t^{i-2}, \dots, h_t^1] \quad (2)$$

Only the last biLSTM layer is used to generate the final sentence representation. Assuming m layers in total, d_m to be the hidden state dimension of the last layer, that is defined as $H_m = (h_1^m, h_2^m, \dots, h_n^m)$ the final sentence representation v is obtained by applying row-max-pooling over the last layer:

$$v = \max(H^m) \quad (3)$$

With each $h_t^m \in \mathbb{R}^{2d_m}$ and $H^m \in \mathbb{R}^{2d_m \times n}$ the resulting sentence vector $v \in \mathbb{R}^{2d_m}$ essentially captures the highest value of each dimension over all timesteps⁷.

⁷ d_m is multiplied by 2 since the biLSTM creates d_m features for going through the sentence forwards and backwards respectively.

2.3.2 Classification

A two-layer MLP using ReLU as activation function and a final softmax-layer is used for the prediction. The input to the classifier m is the concatenation of the sentence representations v_p and v_h for p and h respectively together with the element-wise distance and the elementwise product, denoted as \otimes of both representations:

$$m = [v_p, v_h, |v_p - v_h|, v_p \otimes v_h] \quad (4)$$

Even though a multi-layer MLP theoretically would be able to learn the latter two features, Mou et al. (2015) showed that this particular feature concatenation gives a performance gain for neural models for NLI.

2.3.3 Training

For all our reimplementations using pytorch⁸ of the model we follow the parameters of the original paper of Nie and Bansal (2017). The model is trained using Adam (Kingma and Ba, 2014) parameter optimization, cross-entropy loss as objective function and minibatches of size 32. To avoid overfitting a dropout of 0.1 is applied on each layer of the MLP and the accuracy is evaluated regularly on a different dataset than the train data, the development set, as it is common practice in machine learning. The final performance is estimated by evaluating the best model based on the accuracy on the development set on unseen hold-out data, the test set. 300-dimensional GloVe 840B pretrained word-embeddings (Pennington et al., 2014) are used and finetuned during training. Three additional word-vectors are added, one for unknown words, as well as one to indicate the start and one to indicate the end of a sentence. The learning rate starts with 0.0002 and is reduced by half every second iteration. We conduct our experiments with different re-implementations of this model, partly due to using fewer parameters by reducing the dimensionality of the components, partly due to changes within the original paper.

2.3.4 Residual Encoder and Reimplementation Variants

In a second version of the paper, Nie and Bansal (2017) introduced the Residual Encoder, slightly adapting the way sentences are encoded. In order to create the input to the i th biLSTM layer x_t^i concatenation all previous outputs ($h_t^{i-1}, h_t^{i-2}, \dots, h_t^1$) together with w_t , naturally leads to a tremendous increase of parameters. By using residual connections, instead of concatenating all previous outputs, they are added up, thus equation (2) changes to

$$x_t^i = [w_t, h_t^{i-1} + h_t^{i-2} + \dots + h_t^1] \quad (5)$$

and reduces the parameter size.

Implementation Variants

We use the following implementations of the model. The performance comparison between the models based on SNLI⁹, is listed in Table 1 and do not differ tremendously from what Gong et al. (2017) estimated to be the human performance on the same task.

Model	SNLI train acc.	SNLI dev acc.	SNLI test acc.
Shortcut-Stacked Encoder [†]	87.4%	85.2%	84.8%
Shortcut-Stacked Encoder ^{††}	89.4%	86.0%	85.4%
Residual Encoder [†]	91.1%	85.9%	85.8%
Residual Encoder [◇]	91.0%	87.0%	86.0%
Human Performance (Gong et al., 2017)	-	-	87.7

Table 1: Accuracy in percent of different implementations of the model from Nie and Bansal (2017), achieved on the SNLI dataset compared with human performance.

- We refer to Shortcut-Stacked Encoder[†] as the first re-implementation. This uses 256×2 , 512×2 and 1024×2 dimensions for the three layers of the sentence encoding biLSTM and 1600 dimensions in the classifier MLP.

⁸ <http://pytorch.org/>

⁹ SNLI is a huge dataset for NLI and will be explained in §3.2

-
- We refer to Residual Encoder[†] when we use our own re-implementation with residual connections. The sentence-encoding biLSTMs each have the dimensionality of 600×2 and the layers of the MLP of 800.
 - We refer to Residual Encoder[◇] when we use the final published version of Nie and Bansal (2017) with their provided code¹⁰. This model has the same parameter sizes as Residual Encoder[†].
 - We refer to the plain model name, when talking about the model structure in general.

¹⁰ <https://github.com/easonnie/ResEncoder>

3 Related Work

Much work has been done to create strong models for NLI and we show some successful strategies in section §3.3 and relevant datasets in section 3.2. Before to the excessive usage of neural networks, many models heavily relied on external resources, that have either been manually created in order to improve tools for NLP or arisen for a different purpose, but can also be exploited. In §3.1 we show an overview of some external resources that might improve the performance of neural models on NLI. While most neural models rely only on distributed word-representations as external information and perform quite good, prior work (Bos and Markert, 2005; Tatu and Moldovan, 2005) heavily depended on resources. In §3.4 we show several approaches trying to combine the power of well structured, knowledge-rich resources with the generalization power coming from neural models with distributed word embeddings.

3.1 External Resources

A large variety of knowledge bases exist, containing for instance lexical relations or commonly known world knowledge, which can be helpful for improving of NLI. Research has shown that both, manually created and automatically created resources, can successfully be applied in many tasks of NLP. In this section we only show WordNet and Wikipedia, containing different information that we consider to be useful for NLI and NLU, as well as two resources combining multiple resources and thus providing a huge amount of readily-available knowledge.

3.1.1 WordNet

WordNet (Miller, 1995) is a famous, manually created lexical resource for the English language consisting of three lexica for four different Part of Speech (POS), one for verbs, one for nouns and one for adjectives and adverbs respectively (Jurafsky and Martin, 2008).

Structure of WordNet

Mainly focusing on nouns¹¹ it differentiates between the more frequent class of *common nouns* like “table” and *instances* like “Berlin”. All words are represented by their lemma and due to polysemy contain one or more senses, namely *synsets*. Synsets are the main building blocks within the WordNet ontology, containing a sense description and examples. Figure 3 displays 6 different senses for the lemma “table”. It is noteworthy that the sense of table (as tabular array) greatly differs from the sense as furniture or tablelands while metaphorical senses strongly correlate with the sense of table as a furniture, encoding much more fine-grained sense-differences. While lemmata within the same synset refers to synonymy, other lexical relations like hypernymy, antonymy and holonymy (as described in Section 2.2, however more fine-grained¹²) are defined via labelled links between synsets. Thus, WordNet holds valuable knowledge for detecting lexical inferences in natural language.

- S: (n) **table**, **tabular array** (a set of data arranged in rows and columns) “see table 1”
- S: (n) **table** (a piece of furniture having a smooth flat top that is usually supported by one or more vertical legs) “it was a sturdy table”
- S: (n) **table** (a piece of furniture with tableware for a meal laid out on it) “I reserved a table at my favorite restaurant”
- S: (n) **mesa**, **table** (flat tableland with steep edges) “the tribe was relatively safe on the mesa but they had to descend into the valley for water”
- S: (n) **table** (a company of people assembled at a table for a meal or game) “he entertained the whole table with his witty remarks”
- S: (n) **board**, **table** (food or meals in general) “she sets a fine table”; “room and board”

Figure 3: Example of different synsets of the lemma “table” (only noun senses) within WordNet, taken from <http://wordnetweb.princeton.edu>.

Usage and Issues

When using WordNet in applications one has to identify the correct sense out of many possible synsets for a given lemma. This may be done using proper algorithms for Word Sense Disambiguation (WSD). Another simple and frequently used heuristic is to always choose the first synset, which typically reflects the most common sense (McCarthy et al., 2004). As shown in Figure 3, word-senses are defined with different granularities, sometimes varying only with subtle differences that are not required by most applications. Subsequently this reduces the interpretability of path lengths of lexical relations between two synsets. For instance, identifying that “sunflower” is a hyponym of “plant” requires the traversal over five edges (*sunflower* → *flower* → *angiosperm* → *spermatophyte* → *vascular plant* → *plant*). At the same time, identifying that a “church” is a “building” can be identified by only traversing over two edges (*church* → *place of worship* → *building*) and traversing similarly over five edges leads to the synset “whole, unit”, covering both, living things and objects. This is a known issue (Resnik, 1995) and strategies exist to reduce the complexity of WordNet using sense clustering (Prakash et al., 2007).

¹¹ WordNet 3.0 contains 117,798 nouns, 11,529 verbs, 22,479 adjectives and 4,481 adverbs (Jurafsky and Martin, 2008).

¹² For example, WordNet differentiates between *hypernyms* for common nouns and *instance-hypernyms* for instances, or distinguished between *part-*, *ember-* and *substance-holonyms*.

3.1.2 Wikipedia

While WordNet contains manually annotated lexical relations and is easily automatically accessible, Wikipedia¹³ is a huge multi-lingual, continuously growing encyclopedia, maintained by many volunteers. Also mostly focusing on nouns, due to the nature of containing encyclopedic information, it contains a large variety of factual information about named entities, that many other lexical resources lack (Gurevych et al., 2016). While not being made for the purpose of being a lexical knowledge base, it may be seen as partially annotated due to artifacts like hyperlinks, thus can be interpreted similarly and even accessed using available tools in a programmatic manner (Zesch et al., 2008). Gurevych et al. (2016) describe the following information types that can be exploited to retrieve lexical information:

- The first paragraph of an article can be interpreted as the *sense definition*, since every article covers only one aspect by design.
- *Sense examples* can be retrieved from the context, surrounding a hyperlink that links to the entity of interest, showing how the term is used.
- Hyperlinks between articles can be considered as *sense relations*.
- Due to interlinked articles in different languages, the corresponding titles usually can serve as translation *equivalents*.

Wikipedia has successfully been used in many applications for NLP and even though we do not conduct experiments within this work using Wikipedia, it clearly contains rich factual and world knowledge that can be helpful for NLI systems.

3.1.3 Derived from multiple Knowledge Bases

Yet Another Great Ontology (YAGO) (Suchanek et al., 2007) combines the high coverage of Wikipedia with the clean taxonomy of WordNet, leading to a very knowledge rich resource. While YAGO mainly targets world-knowledge with Wikipedia being tremendously larger than WordNet and containing relations to express facts derived from it, UBY (Gurevych et al., 2012) aims for lexical semantic richness. In addition to Wikipedia and WordNet, seven other resources are combined together providing lexical semantic knowledge in German and English by so called *sense axis*, realtions between two senses of different lexicons. UBY provides an easy-to-use API, making its high coverage knowledge programatically accessible to NLP applications. Having these knowledge-rich resources available, but for the most part decoupled from neural models, still lacking this knowledge, stresses the benefit of combining these two worlds.

3.2 Datasets for NLI

As neural models usually require a huge amount of data for their training, they were not successfully applicable to NLI tasks until the release of The Stanford Natural Language Inference Corpus (SNLI), reaching state-of-the-art results. Previous NLI tasks like FraCas (Cooper et al., 1996) or the PASCAL challenge (Dagan et al., 2006) only had very limited amount of training data, such that neural models could not be used. Some datasets like Sentences Involving Compositional Knowledge (SICK) (Marelli et al., 2014) or the Denotation Graph entailment set (Young et al., 2014) increased the amount of samples at the expense of using artificially created sentences and/or automatically labeling, reducing the textual quality and adding noise. Since the focus in this work is on neural models, only the relevant datasets for this purpose are introduced.

3.2.1 SNLI

With the release of SNLI (Bowman et al., 2015) researchers were able to apply neural models for the task of NLI using distributed word and phrase representations. The corpus consists of 570,152 human written sentence pairs and differentiates between the labels *entailment*, *neutral* and *contradiction*.

Event co-reference

A drawback of all previously existing resources for NLI that is handled by Bowman et al. (2015), is the fact that even humans may assign different labels to a sentence pair, based on their subjective interpretation of a sentence, that all can be valid. This issue can be demonstrated using the following sentence pair:

Premise: *Young people are demonstrating in San Francisco.*
Hypothesis: *Young people are demonstrating in New York.*

¹³ <https://www.wikipedia.org/>

One could clearly argue the sentence-pair should be labelled as *neutral*, since there could be people demonstrating in both towns. However it is also legitimate to interpret these as contradicting sentences, if one considers both sentences to be describing the same event. While both sentences may be true when describing different potential scenarios, only one of them can be true if they refer to the same. In order to reduce noise coming from these inconsistent interpretations, the labeling scheme within SNLI must be fixed beforehand. Specifically they choose the labelling scheme to be based on event-coreference, the latter of the two explanations, as otherwise only very general statements would result in *contradiction*.

Generation

In order to create SNLI Bowman et al. (2015) used image captions from the Flickr30k corpus (Young et al., 2014) as premises and let human workers create according hypothesis for each label respectively using Amazon Mechanical Turk by asking them to write alternative captions that

- definitely also are a true description of the photo (**entailment**)
- might be a true description of the photo (**neutral**)
- definitely are a false description of the photo (**contradiction**)

The workers only saw the image caption, not the image itself, but were encouraged to use common world knowledge, enabling the creation of inferences that require additional information of the world, that is not available in word-embeddings¹⁴. While this process simplifies the task of assuming event-coreference, the sentences within SNLI are rather simple and short, due to the nature of image captions.

Looking into data

As we conduct most of the experiments of this work on SNLI it is important to get an understanding how sentences in this dataset look like. As previously mentioned the vast majority of sentences are rather simple and might even be phrases only rather than proper sentences due to omission of a verb. In addition to that, sentences might be written in proper English, but also might contain spelling or punctuation errors, be lowercased only or describe highly unrealistic scenarios. Table 2 shows selected sample sentence-pairs, taken from the SNLI dataset.

Premise	Hypothesis	Label
(1) The large brown dog jumps into a pond.	The dog is getting wet.	<i>entailment</i>
	The dog is a chocolate Labrador Retriever.	<i>neutral</i>
	A white cat is sunning itself on a windowsill.	<i>contradiction</i>
(2) A woman is handing out fruit.	A woman is passing out different types of fruits.	<i>entailment</i>
	A woman is handing out oranges.	<i>neutral</i>
	A fruit is handing out a woman.	<i>contradiction</i>
(3) A basketball game.	A sports game.	<i>entailment</i>
	A basketball game between rivals.	<i>neutral</i>
	A volleyball game.	<i>contradiction</i>

Table 2: Example sentence pairs, taken from SNLI, showing typical sentences within the dataset.

The first column displays the premise, the original image caption, in the second column three hypothesis are shown, created by the workers for each label respectively. Several characteristics of the dataset and types of required knowledge to solve the task can be seen here. The first examples (1) require the model to have some factual knowledge that a *Labrador Retriever* is some kind of *dog*, and *chocolate* is paraphrasing *brown*. Since *Labrador Retriever* is a possible substitute for *dog* but more specific, the sample is labelled as *neutral*. The according entailing hypothesis requires an even deeper understanding of the world, as the system needs to know that a *pond* is filled with water and anything that goes into water is *getting wet*. The contradicting sample shows two frequently occurring characteristics. Not only has *dog* been replaced by *cat*, but also the color and the activity changed. We found that in many contradicting hypothesis several contradicting words with respect to the premise exist, obviously making the task easier, as it would be sufficient

¹⁴ For instance (taken from SNLI) *snow* is paraphrased as *frozen particles of water* and requires very deep factual knowledge to be understood correctly.

to only detect one indicator. Additionally it has been shown that the creation process of the hypothesis followed some unconscious heuristics of the worker (Gururangan et al., 2018). Specifically the replacement of *dog* to *cat* occurs often enough, that the presence of *cat* in the hypothesis alone is a strong indicator for contradiction already.

The sentences of the second example (2) are based on paraphrasing, representing the same meaning, (entailment), are more specific term in the hypothesis as in the previous example (neutral) and semantic role reversal (contradiction), which is somewhat interesting, as it requires to model to leverage word order, while a simple Bag of Words (BoW) approach would fail here.

In contrast to (1) the sentences in (3) only require very shallow knowledge. Here, the word *basketball* is substituted by its hypernym¹⁵ *sport*, thus still covering the original meaning by being more general. The next sentence gives some plausible additional information not given in the premise, hence neutral. In the last contradicting sentence the model has identify that *basketball* and *volleyball* are mutually exclusive, as generally one can only play on of the two sports at a time. As both terms are hyponyms of the term *sport* we refer to them as *co-hyponyms*. Amongst other relations like antonyms, co-hyponyms are especially challenging for models using distributed word representations and will be a central point in section 5. While sentences are often longer than in this sample, the required knowledge specified in (3) is most present within SNLI.

3.2.2 MultiNLI

SNLI received some criticism within the research community (Chatzikyriakidis et al., 2017; Williams et al., 2017), mainly due to its simplicity, coming from the fact, that all premises are taken from a single genre only, namely image captions. Thus, SNLI is very limited to only visual scenes, neglecting many other aspects like temporal reasoning, modality or belief. Williams et al. (2017) introduced with MultiGenre Natural Language Inference Corpus (MultiNLI) a new dataset, overcoming these drawbacks.

Generation of MultiNLI

The authors followed the same generation procedure as has been done by Bowman et al. (2015), but instead of relying on image captions only, they took into considerations other genres from Open American National Corpus (OANC)¹⁶ (Ide and Macleod, 2001; Ide and Suderman, 2004, 2006) as well as several freely available fiction work, resulting in 10 additional genres with 392,702 new sentence pairs for training and 20,000 for development and test respectively. A major motivation for the creation of MultiNLI was to put more emphasis on the role of NLI as evaluation benchmark of NLU that SNLI failed to provide due to its narrow coverage. Therefore only five of the new genres are present within the train data, while the remaining five genres only occur in the testing sets and serve for cross-domain transfer learning and domain adaption. The performance on this dataset is measured in two figures, *matched* examples are derived from the same source as training samples, while *mismatched* examples differ from those seen during the training. This motivation becomes also clear from the corresponding Shared Task (Nangia et al., 2017), allowing any kind of external resources (including the ones that were used to derive the premises) but only accepting sentence-encoding models¹⁷ to evaluate sentence representations learning with respect to NLU. MultiNLI has been shown to be harder than SNLI (Williams et al., 2017), the best performing model of the RepEval 2017 Shared Task reaches 74.9% matched and mismatched accuracy (Chen et al., 2017c) using ensembles and 74.5% matched, 73.5% mismatched accuracy using a single model (Nie and Bansal, 2017).

Looking into data

Table 3 depicts a few samples of different genres¹⁸. One can see how different genres broaden the scope of language that is used to express inferences. A system needs to deal with temporal information and less visualizable terms like *appreciate* or *benefit*.

As the authors followed the guidelines used for SNLI and also assume event-coreference, both datasets are highly compatible, only differing in the range of genres and thus diversity of language. In fact, MultiNLI is even distributed in the same data format and a common practice is to include data from SNLI when training models for MultiNLI (Nie and Bansal, 2017; Balazs et al., 2017; Yang et al., 2017).

¹⁵ At least in one sense. Hypernymy refers to the *is-a* relationship meaning *basketball*(hyponym) *is a* *sport*(hypernym). Word-relations are discussed in more detail in section 3.

¹⁶ Genres from OANC: *Government, Slate, Telephone Speech, Travel Guides, 9/11 Report, Face-to-face Speech, Letters, Nonfiction Books, Magazine*

¹⁷ These models encode each sentence individually and are explained in §3.3.

¹⁸ Taken from <https://repeval2017.github.io/shared/>

Premise	Hypothesis	Label	Genre
The Old One always comforted Ca'daan, except today.	Ca'daan knew the Old One very well.	<i>neutral</i>	Fiction
Your gift is appreciated by each and every student who will benefit from your generosity.	Hundreds of students will benefit from your generosity.	<i>neutral</i>	Letters
At the other end of Pennsylvania Avenue, people began to line up for a White House tour.	People formed a line at the end of Pennsylvania Avenue.	<i>contradiction</i>	9/11 Report

Table 3: Example sentence pairs from MultiNLI, taken from RepEval 2017 Shared Task, showing samples of different genres.

3.2.3 SciTail

SciTail (Khot et al., 2018) is yet another dataset for NLI, designed to address a different problem of previously existing datasets¹⁹. The targeted problem is of previous work is, that either the premise or the hypothesis was specifically for this task created, thus neglecting the kind of naturally occurring inference problems of any endtask like QA. It is comparably smaller, consisting of only 27,026 examples and only distinguishes between two labels, *entailment* and *neutral*. *Entailment* is defined as in SNLI and MultiNLI, saying that the premise supports the hypothesis. All cases where the hypothesis is not supported by the premise however are classified *neutral*.

Generation of SciTail

In order to retrieve premise and hypothesis from a resource rather than creating one sentence for the specific purpose of NLI, Khot et al. (2018) took a different approach to generate the corpus. The dataset originates from school-level multiple choice questions for science QA (Welbl et al., 2017). Those questions generally require sophisticated reasoning capabilities in order to answer them correctly.

1. Given the short factual answer candidates, a new sentence was syntetized using the question and correct answer, that serves as the hypothesis. For instance the question “When waves of two different frequencies interfere, *what phenomenon occurs?*” and the answer “beating” is transformed into “When waves of two different frequencies interfere, *beating occurs*” (Khot et al., 2018).
2. A large background corpus with relevant information from Clark et al. (2016) was used to generate candidate knowledge sentences for each question using Information Retrieval (IR) for the premise.
3. While hypothesis, derived from an incorrect answer, can be assumed to be not-supported by the premise, those derived from a correct answer are not necessarily supported by the sentence gained from the background corpus. Thus, samples were crowdsourced annotated to ensure a correct labelling, only keeping those samples as *entailment*, that were labelled to have *Complete Support*²⁰.

Comparison with SNLI and MultiNLI

Due to its design, SciTail is different in nature to the two previous datasets. Neither does it contain contradicting examples nor does it assume event-coreference, as sentence-pairs in this dataset are more based on factual information. Table 4 shows sample sentences of the SciTail dataset. Clearly all of them contain factual information, whereas in the previous shown datasets, sentences tend to be more situational. The premise can be relevant for the entailment relation, yet must not be. In both cases the keywords *joint* or *range* appears in all premises, retrieved for both hypothesis respectively.

Due to its relatedness with Scientific QA, the authors claim, that a model reaching a good performance on this dataset for NLI will also score well on an according QA task, as similar NLU is needed.

¹⁹ Ignoring small-scale datasets with less than 1,000 samples.

²⁰ Annotators could decide between *Complete support* (labelled as entailment), *Partial Support* (ignored) and *Unrelated* (labelled as neutral).

Premise	Hypothesis	Label
Bones come together to form joints, most of which are in constant motion.	Joints are the location where bones come together.	<i>entailment</i>
Bone, Joint, and Muscle Disorders Chapter 54 Charcot's Joints Charcot's joints (neuropathic joint disease) results from nerve damage that impairs a person's ability to perceive pain coming from a joint;	Joints are the location where bones come together.	<i>neutral</i>
The time to travel the horizontal distance (the range) is equal to twice the time to reach the peak (maximum height).	Range is the maximum horizontal distance traveled by a projectile.	<i>entailment</i>
First, finding the launch angle for maximum horizontal range in idealized projectile motion.	Range is the maximum horizontal distance traveled by a projectile.	<i>neutral</i>

Table 4: Example sentence pairs from SciTail Task, different premises retrieved for two hypothesis.

3.3 Neural Models for NLI

We follow the SNLI leaderboard²¹ by differentiating between *sentence-encoding* and *inter-sentence-attention* based models, in the following we show an overview about relevant approaches of both areas. The Residual Encoder or Shortcut-Stacked Encoder introduced in section 2.3 belongs to the former class of models.

3.3.1 Sentence Encoding Models

Sentence-encoding models follow the Siamese Architecture (Bromley et al., 1994), meaning they encode both, sentences p and h , individually, with parameters being tied between both sentence encoders. The entailment classification is predicted by a following classifier like a MLP. Doing so, the models put more emphasis on a meaningful sentence representation with the motivation of being more generally applicable and less focused on the specific characteristics of the task at hand (Bowman et al., 2016). Many different strategies are used to create meaningful sentence representations within this class of neural models. This is for instance done by exploiting syntactical information using neural Shift-Reduce-Parsers that creates a linear sequential structure from tree-structures sentence representations (Bowman et al., 2016) or by adding and external memory with read and write operations, capturing the temporal and hierarchical information within natural language (Munkhdalai and Yu, 2017).

Inner-Attention-based Models

Following the intuition that humans only remember certain parts of a sentence after reading it, Chen et al. (2017c) modeled this behaviour using gated intra-sentence attention, by generating the sentence representation using pooling²² strategies over the outputs of the encoding biLSTM, reweighted using the attention gates. The idea of using inner-attention mechanisms is also used by the best performing sentence encoders for SNLI, at the time of this writing reaching 86.3% accuracy (Shen et al., 2018; Im and Cho, 2017). Shen et al. (2018) create sentence representations using the combination of hard and soft self-attention²³. While hard-attention forces the model to only focus on relevant elements of the input sequence, disregarding all others, it is not fully differentiable and thus inefficient to train. Soft-attention methods on the other hand are fully differentiable and weights each element of the input sequence according to their relevance. However by also giving positive, non-zero weights to irrelevant elements it diminishes the emphasis on truly important ones. By first applying hard-attention to retrieve a subset of context-aware elements that is afterwards processed using soft-attention, Shen et al. (2018) leverage the mentioned advantages both techniques. Inner attention is also used by Im and Cho (2017), however their model additionally uses directional masks, that prevent the network from considering following or preceding words in the attention process respectively and distance masks that reduce the attention weight if words are further away to each other. They show that their model outperforms others, especially with longer input sentences, as the result of considering word distance and positional information.

²¹ <https://nlp.stanford.edu/projects/snli/>

²² As done with max-pooling by Nie and Bansal (2017).

²³ A plain attention function calculates the alignment for an input sequence $x = [x_1, x_2, \dots, x_n]$ given a query q . In the special case of self-attention, q arises from the input sequence x itself (Shen et al., 2018).

3.3.2 Inter-Sentence-Attention-based Models

Rocktäschel et al. (2015) showed that models perform significantly better when looking at both sentences simultaneously. This is motivated by the way humans would solve the task of NLI by first reading the premise and creating the understanding of the hypothesis based on the previously read sentence. Since this seems to be superior in the SNLI, many works followed this approach reaching state-of-the-art results. Also in this class of methods memory networks, accessible via attention were applied (Cheng et al., 2016).

Inter-Sentence-Attention-based models used within this work

Parikh et al. (2016) provide a simple network structure called Decomposable Attention, using the assumption that only parts of a sentence are needed for the entailment relationship. They do so by fragmenting the input sentences into subphrases and align fragments of both sentences with each other using attention. Even though they represent sentences in a BoW manner, they reach a remarkable performance. After comparing the aligned phrase pairs, the final sentence representation is retrieved by simple summation over the comparison-vectors, thus reducing without relying on any LSTM-based method and reducing computational complexity. Enhanced Sequential Inference Model (ESIM) (Chen et al., 2017b) is another simple yet strong model, essentially consisting of three different steps. First, words are encoded using biLSTMs such that they represent the context as well as the word itself. Next, similarly to Parikh et al. (2016), they calculate the local inference between elements in both sentences by reweighting the sentence representation based on the normalized attention weights and enhance this information using the feature concatenation of Mou et al. (2015), as done in the Shortcut-Stacked Encoder, however in this approach word order information is preserved by the network in contrast to Decomposable Attention. The final sentence representation is created using pooling²⁴ on the output of a biLSTM, composing the local inference information from the previous step, and is then fed into a classifier. While Decomposable Attention and ESIM achieve competitive results on SNLI we conduct experiments using both models Section 5, showing that these results are rather a matter of memorization than generalization.

Attempt to incorporate WordNet

Very recently Chen et al. (2017a) introduced with Knowledge-based Inference Model (KIM) a neural model incorporating information from WordNet, which at the time of this writing is the single best performing model on SNLI. In their approach, they map WordNet relations, as in Section 2.2, to a real number $0 \leq r \leq 1$, quantifying the relations between word within p and h based on their path length and represent each word-pair with this additional feature vector. This way they create new feature vectors yexternal resources only by a small margin, ranging from 0.1 to 0.6 points in accuracy. Chen et al. (2017a) show that adding WordNet is helpful if less train data is available, however only show limited prove, that the model leverages from WordNet fused relations for the overall improvement in accuracy. In this paper we show that performance on SNLI is not sufficient evidence for the capability of dealing with simple lexical inferences as inferred from WordNet, which suggests that further investigations should be conducted in this direction.

Benchmark

To this date the best single sentence-encoding models on SNLI reach 88.6% (Chen et al., 2017a) ensembles reach up to 89.3% (Tay et al., 2017; Peters et al., 2018; Ghaeini et al., 2018) giving an advantage of 2.3% or 3.0% respectively over the best sentence-encoding model (Im and Cho, 2017). Considering that the estimated human performance on SNLI only is estimated to be 87.7% (Gong et al., 2017) indicates, that research started to slightly overfit on the dataset already.

3.4 Integration of external Resources into Neural Networks

There have been several approaches to integrate knowledge of different kind (as described in Section §3.2) into neural networks. Hu et al. (2016) infer external knowledge represented in logical form using a student-teacher setup, with the teacher being a neural network constrained by the rules acquired from an external resource, the student being a neural network considering the true labels and the constrained predictions of the teacher. By simultaneous training both networks influenced by each others predictions, the integrated within the networks, weighted by their learned relevance. Most attempts however target improving word representations.

3.4.1 Improving word-embeddings

A very intuitive way to integrate external resources is to enrich word-embeddings with additional information. As most neural models depend on those anyway, any improvement of word-representations can be adapted with very limited effort.

²⁴ As opposed to summation in Decomposable Attention. Chen et al. (2017b) evaluate in their experiments, that pooling leads to superior results than summation, due to being less sensitive to the sentence length.

Joint learning of distributional embeddings with external information

Xu et al. (2014) differentiate between *categorical* (attributes of words like “location”) and *relational* (relations between words like “child-of”) knowledge and train the embeddings from scratch using three objective functions simultaneously. They use skip-gram to encode distributional properties, minimize the distance between words sharing the same category and represent a relation as a vector r , such that for a word w_1 , connected to another word w_2 via relation r the equation $w_1 + r \simeq w_2$ holds (evaluated on analogical reasoning, word similarity, topic prediction). Liu et al. (2015) construct enriched embeddings using a constrained optimization problem. Specifically, they create constraints by ranking word similarities such that for instance synonyms should be more similar than antonyms or hyponyms should be more similar to close hypernyms than to distant hypernyms and include these into the training process with skip-gram.

Post-processing existing representations

Faruqui et al. (2015) propose a method called *Retrofitting* (evaluated on lexical semantic evaluation tasks), a post-processing method that can be applied on any pre-trained word-representations. They reduce the euclidean distance between words that are connected with a lexical semantic relation within a resource, while also keeping the representations close to the original neighbouring word-representations. Attract-Repel (Mrkšić et al., 2017) is another retrofitting method, essentially pulling synonyms closer to each other while pushing antonyms apart in vector space while keeping the original distributional information. Similarly Vulić and Mrkšić (2017) build on attract-repel adding hypernym relations for the context of lexical entailment by using an asymmetric distance measure between hypernym-hyponym pairs.

Effectiveness of improved embeddings

The demand of integrating lexical resources such as WordNet has mainly been targeted by enriching word-representations, with previously mentioned approaches being just a small selection. The improvement of previously explained approaches however is either demonstrated by visualizing word-relation vectors, that may not even be exploited by end-to-end neural networks (Levy et al., 2015), or based on evaluations on very low-level tasks like Word-Similarity, Syntactic Relations or Analogical Reasoning, or solely provide intrinsic evaluation. Neural networks for higher level tasks like NLI however reach state-of-the-art performances still relying only on standard pre-trained distributional word-representations like GloVe, even though alternatives exist, and preliminary experiments²⁵ of using enriched embeddings for SNLI have shown no success. We evaluate the possibility of adding enriched embeddings following the idea of Rücklé et al. (2018), that different embeddings encode complementary information and thus concatenating them. However we focus our experiments on the integration of knowledge on a higher level, the sentence representation, due to limited reported success on end tasks using richer embeddings, though many of those embeddings exist.

²⁵ These experiments have been conducted by Vered Shwartz in prior work and are not part of this work.

4 Understanding Shortcut-Stacked-Encoder

In this section we give analyse the sentence representations of Shortcut-Stacked Encoder[†] by visualizing how they encode (Section §4.2) and leverage (Section §4.3) information from natural language text, coming from SNLI. Additionally we show experiments, underlining the presented insights.

4.1 Motivation

The major downside of neural networks is the lack of interpretability (Goldberg, 2017), thus their capabilities on a lower level can only be estimated by finding meaningful evidence for their failures or successes on the task at hand. While analysing errors may lead to conclusions *what* does not work, *why* it does not work is in many cases left to intuition. Other machine-learning classes like probabilistic or symbolic techniques do not suffer from this problem, leading to an increasing interest in visualization techniques for neural networks. Most visualisations of sentence-representations to date focus on attention-based approaches showing how words are aligned to each other such as by Shen et al. (2018) or Im and Cho (2017). To the best of our knowledge, no insights have been gained to understand the final sentence representation in vector space. In this section we demonstrate how this representation, arising from max-pooling, can be interpreted, using the Shortcut-Stacked Encoder[†] as the model to analyze. Intuitively, understanding how the Shortcut-Stacked Encoder[†] encodes information can be helpful for the task at hand of improving it using external resources.

While we did not manage to leverage the insights gained in this chapter to increase the performance, it might be helpful for future work. NN nicht gut interpretierbar

4.2 Insights on the sentence representation

In this section we show how we analyse the information that is present within the sentence representations, what kind of information is encoded and demonstrate, that the sentence representation can manually be adjusted in a meaningful way.

4.2.1 Approach

We use Shortcut-Stacked Encoder[†], trained on SNLI, for our analyses. This model creates for input each sentence x , consisting of words, represented as x_i , a sentence representation $r \in \mathbb{R}^{2048}$ with r_j being the j th dimension of r . Arising from x , r captures the relevant information for the task at hand and is used in many neural networks without a deeper understanding what each r_j actually encodes. We shed light into the dimensionewise meaning of the sentence representation by identifying which word is responsible for the actual value of r_j .

Method

For simplicity, We explain our applied method and the reason why we use Shortcut-Stacked Encoder using a more general neural architecture of LSTMs, a simple uni-directional RNN. Figure 4 (left) shows the recursive workflow of such a RNN, following the notations of Goldberg (2017). Maintaining an internal state $s \in \mathbb{R}^m$, for m -dimensional representations, the

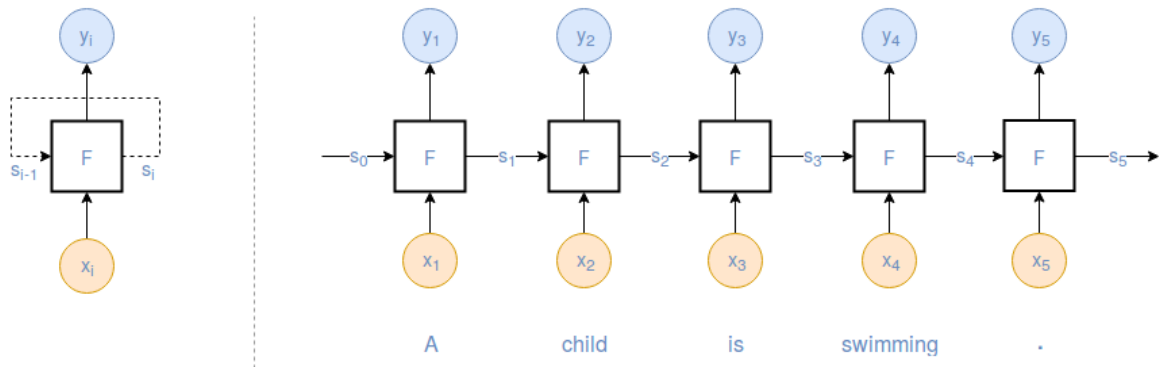


Figure 4: General architecture of a RNN (left). Example sentence in an unrolled RNN (right).

network recursively iterates over the input sequence x , aggregating in each timestep the previous state $s_{i-1} \in \mathbb{R}^m$ with the current input x_i using the function F . This state is then used for the next iteration and output via a mapping function

as $y_i \in \mathbb{R}^m$. Multiple implementation variants exist of F and what is shared across iterations. LSTMs for instance use several neural gates to learn what information should be used, output or forgotten. This procedure can be seen with an example sentence by unrolling the network in Figure 4 (right). In typical setups a neural network may either choose to use s_t or y_t for a sequence length of t as the final sentence representation (Goldberg, 2017), since the network iterated over the full input sequence and contains the relevant information, if optimized for it. Even though the architecture of different versions of RNN may be well understood and has a logical meaning, the actual procedure of deriving concrete representations within a trained model is hard to understand. We leverage the fact that the Shortcut Stacked Encoder uses max-pooling over all y_i to gather the sentence representation rather than using y_t or s_t by identifying what y_t has the highest value within a given dimension and mapping this dimension to the word x_t of the input sentence. As an example consider the sentence in Figure 4 (right). For each timestep t a new vector y_t is produced. As done by Nie and Bansal (2017) we concatenate all y_t to a matrix $\mathbb{R}^{m \times t}$, with m being the representation size and each vector y_t being the t th row within M . Assuming a dimensionality of $m = 3$, an exemplary matrix M for the given sentence “A child is swimming.” is displayed in Figure 5. Additionally to creating the sentence representation r by applying word-wise

$$M = \begin{bmatrix} 1 & 4 & 7 & 2 & 0 \\ 2 & 9 & 4 & 1 & 1 \\ 0 & 3 & 2 & 8 & 2 \end{bmatrix} \xrightarrow{\text{argmax}} r = \begin{bmatrix} 7 \\ 9 \\ 8 \end{bmatrix} a = \begin{bmatrix} 3 \\ 2 \\ 4 \end{bmatrix} \xrightarrow{\text{map}} \begin{bmatrix} \text{is} \\ \text{child} \\ \text{swimming} \end{bmatrix}$$

Figure 5: Visualized example of extracting interpretable information of the max-pooled sentence representations with a dimensionality of 3.

max-pooling on M , we collect the vector a , containing the column indices, that are responsible for the values within r . These can directly be mapped to the word of the source sentence and thus be interpreted by humans. It should be noted that due to the nature of the multi layer biLSTM each y_t does not only contain the word at x_t but its context. While this somehow may lead to less accurate mapping, we found that the chosen method is sufficient to gain some meaningful insights on sentence encoding.

Analysed data

To reduce noise and aiming for sentences that Shortcut-Stacked Encoder[†] seems to have a proper understanding about, we sample 1000 sentence representations from the SNLI train data in the following strategy. We group all sentence pairs (p, h) sharing the same premise and only keep groups if all samples belonging to the same group are classified correctly. Thus, we reduce the amount of sentences that are definitely misunderstood by the model, that would be harder to interpret. For now we are not interested in the actual relation between p and h and therefore create a pool of the remaining sentences, by treating p and h equally and splitting their connections apart. After removing duplicate sentences, the most frequent sentence length for the remaining representations is 8. To reduce noise that may arise from different sentence lengths, we only consider sentences of a length of 8 and randomly sample 1000 sentence representations. All experiments in this chapter are based on the same instances, unless otherwise stated.

In addition to the representation values each sample contains the following information:

- **Token:** The tokens that triggered the maximum value for the representation.
- **Token position:** Positional information about the responsible tokens within the sentence.
- **Lemma:** The lemmata of the responsible tokens.
- **POS:** The POS tags of the responsible tokens.
- **Dependency Parse:** The tags of the responsible tokens within the dependency parse tree.

Lemmatizing, POS-Tagging and dependency parsing were conducted using spaCy²⁶.

4.2.2 Detection of relevant dimensions

As commonly done when analysing data we start by showing a rough look into the sentence representations at hand. Typically, the Standard Deviation (SD) within a dimension correlates with the with the relevance for decision making.

²⁶ <https://spacy.io/>

Naturally, a dimension that does not change its value and thus being close to a constant is not informative, while a value with a high SD can be considered informative (Bishop, 2007). We calculate SD over all dimensions, depicted as a histogram in Figure 6. We plot the standard deviations in a discrete space using a bin size 0.05. For each of the 2048

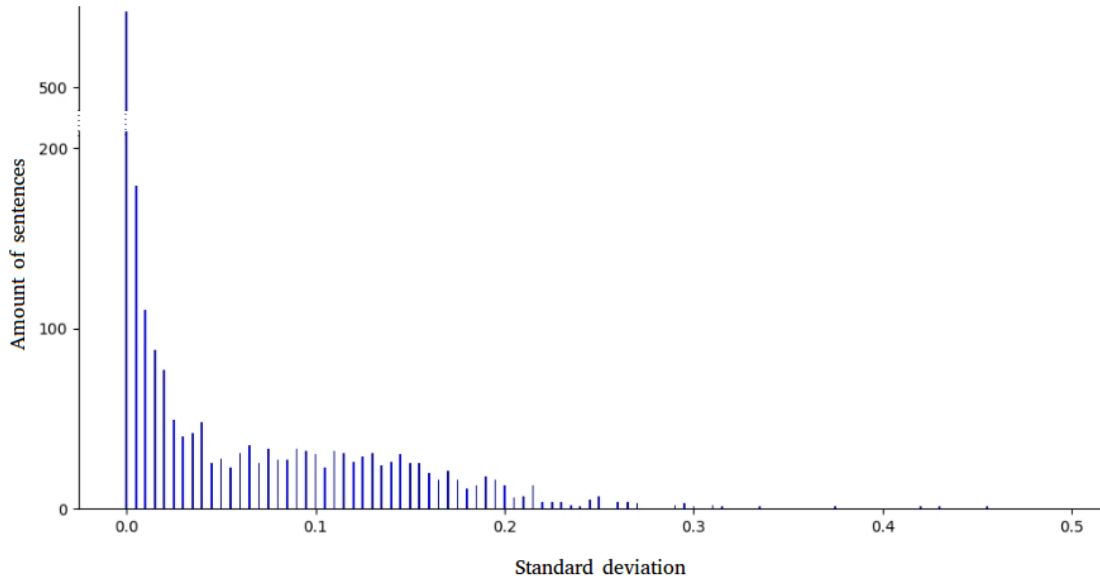


Figure 6: The standard deviation within a dimension of sentence representations (x-axis) by the amount of dimensions with the given standard deviation.

dimensions we calculate its SD to assign them to the correct bin. The amount of dimensions with the given SD is shown on the y-axis, note that the upper part of the plot is truncated for the sake of compactness. As can be seen, only a very tiny fraction of the dimension shows a large variation, the vast majority contains more or less the same value, regardless of the sentence. This obviously does not mean, they contain no information at all, as they may only be used to encode information that is rarely present within the data, however it serves as a reliable source, what dimensions are relevant to the model.

A naive approach to identify dimensional encoding

An intuitive approach to identify, what is encoded within the sentence representation, is to find common similarities between the words across all sentences, that are responsible for the according dimension. Especially the task of NLI we assume *semantic*, *syntactic* or *positional* information to be required. Those can all be inferred using the features we extracted in Section §4.2.1. Similarities between words heavily depend on the context they appear in (Dagan, 2000). For instance one could consider a car and an identical reconstruction in original size of the same car as similar, whereas a horse is very distinct. Adding additional information that one needs to reach a destination in short time, he or she is more likely to consider the horse similar to the car, deciding between these two options. This essentially comes to a major problem when investigating semantic encoding without prior knowledge of what attributes may be considered relevant. We therefore investigate the sentence representation using excessive manual search in a top down manner, by first searching for patterns across all dimensions. In Section §4.2.3 we will look into some dimensions in detail.

A tool for sentence representation visualization

In order to evaluate many patterns with minimal time effort, we create a visualization tool, capable of dynamically generating any labelling scheme for responsible words based on the features described previously. A sample visualization is shown in Figure 7. This grid-plot visualizes for each row the responsible words for one dimension, listed on the right side as (<rank in terms of SD²⁷> - <dimension index>), colored based on the attributes of interest. In this particular case words are colored by their position within the sentence. Each column refers to the same sentence along different dimensions. As a trade-off between explanatory power and clarity we always plot 300 sentences on 300 dimensions, which are either ordered by SD or already pre-sorted by the frequency²⁸ of a label of interest. In this particular case,

²⁷ All dimensions are ranked by their SD, giving an intuition of the expressiveness of the dimension.

²⁸ Even though we only use 300 sentences and dimensions for plotting, calculations are based on all the selected data.

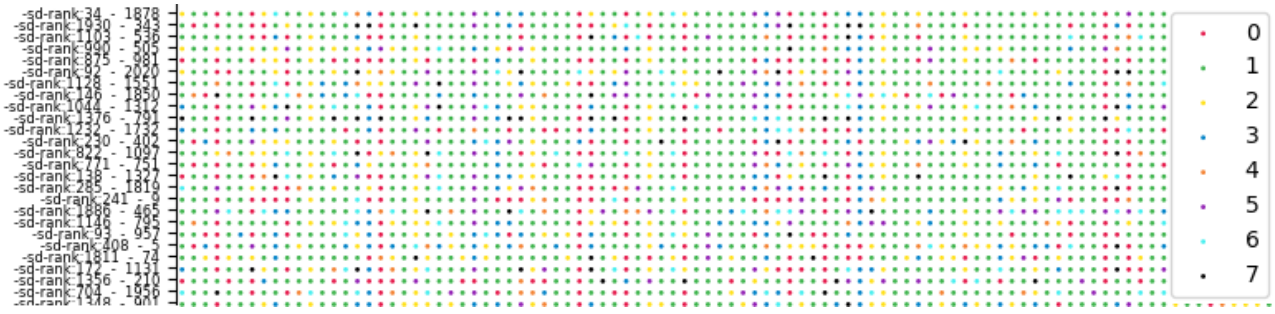


Figure 7: An extraction of a grid-plot, showing dimensions with the position within the sentence of the word, responsible for the dimensional value.

dimensions are ordered by their frequency of words on the $\text{uposition } 1$, meaning the upmost dimension received its values from the second word (index 1) more than any other dimension. Looking for patterns across many sentences, we focus on horizontal lines with the same coloring or colors referring to attributes that may be interpreted similarly. Vertical lines indicate different differences across sentences with respect to the attribute of interest.

Interpretation of positional Information

Word-ordering is crucial with respect to the meaning of a sentence. We evaluate if certain dimensions are aligned to specific word positions, only meant to encode the meaning of the word at this place. Figure 7 shows dimensions that are heavily influenced by the second word, indicated by the vast majority of green dots. And indeed, several also very informative dimensions are dominated by the second word of a sentence with some noise, primarily stemming from other word positions from the beginning of the sentence. Considering the nature of sentences of SNLI, as presented in Section §3.2.1, this is however not enough evidence to conclude, those dimensions correspond (solely) to positional information within the sentence. Consisting merely of simple sentence structures of even only phrases it is very likely that the second word corresponds to a noun, most likely describing the main aspect of the image. Optional preceding articles or adjectives may cause this noun to have varying positions between one and three. Looking at the coloring of vertical lines this assumption is backed, as for each sentence, the responsible word arises fairly consistently from the same position across dimensions, indicating that this stems from the encoded attribute rather than noise. In general we find no meaningful²⁹ dimensions encoding solely positional information, either absolute or relative, without being correlated strongly with another attribute.

Finding syntactic dimensions

This warrants more investigation using a different labeling scheme, and we look for clues based on the POS tags. Tokens are labelled using the Penn Treebank Part-of-Speech Tagset (Marcus et al., 1993) and present in our data with the originally assigned labels. Figure 8 shows an extract of dimensions labelled by POS tags, pre-sorted such that dimensions with any single dominant label are shown first. We aggregate different POS-tags referring to the same concept together, thus for instance all nouns NN, NNS, NNP and NNPS are all labelled as NN. Several patterns can be seen within this plot especially punctuation seems very well presented at first sight (green and orange). Looking into the actual data and considering their very low SD, these dimensions seem in fact less important. We observe similar issues for dimensions that are dominated by articles (orange). More interestingly are nouns (yellow), being very dominant in diverse dimensions, including dimension 1878 (fourth row), which also was well represented by the second word when checking for positional information (first row in Figure 7). This supports our intuition that word positions are not directly encoded but merely correlated to other features, like an early noun in this case. Also verbs (blue) are well presented in two dimensions. This plot suggests that indeed that model captures syntactic information to some extent, represented within the according dimensions respectively, however also shows some drawbacks of our initial naive approach:

1. **Correlation:** As we have shown different attributes may correlate with each other, thus it is unclear if a found pattern is merely a side product of a correlated feature or the main thing being encoded by a dimension.
2. **Non existent information:** Interpreting the meaning of a dimension by the responsible word is only possible if this word indeed characterizes the dimension. In case of positional information we can be certain, that each sentence contains a word that reflects the information with respect to our labeling scheme, as every sentence contains

²⁹ We do find several dimensions that mostly arise from the first word, yet the resulting value is mostly constant across all sentences and stems mainly from articles.

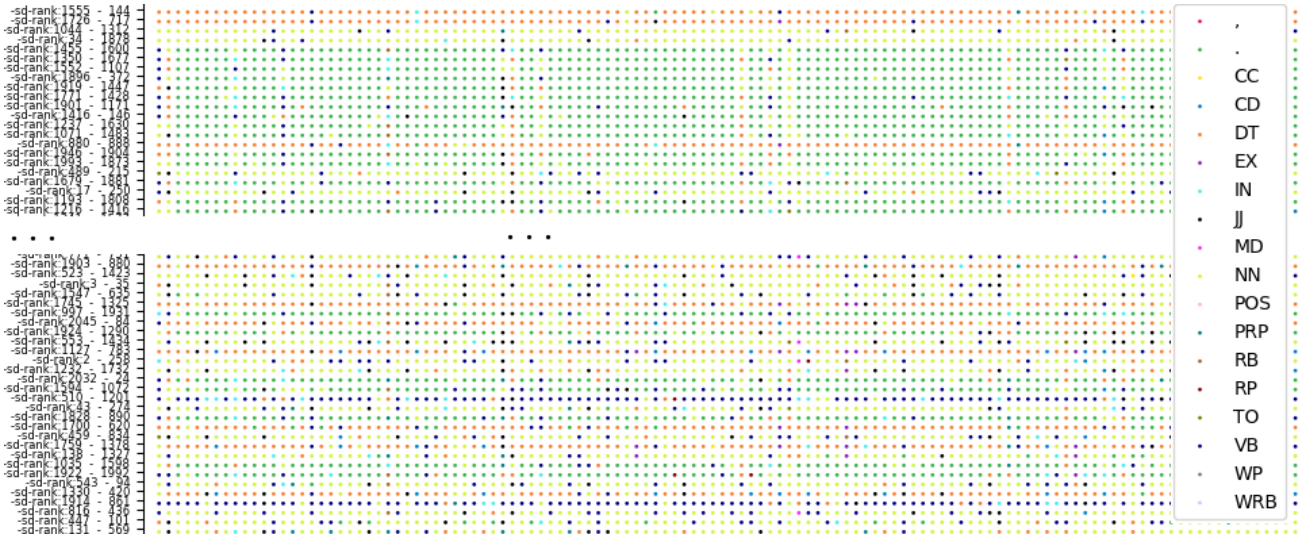


Figure 8: An extraction of a grid-plot, showing syntactical information using the POS tag with pre-sorted rows to have a single dominant label.

words with all positions. However when looking for encoded information that is not present amongs all sentences, another arbitrary word will still be responsible for the dimension’s value. In our caase when finding syntactical clue for instance, most sentences have punctuation, nouns, determiners or verbs. A dimension representing adjectives however will always look very noisy, as it can only be represented by adjectives in sentences containing one.

3. **Representation value:** We did not yet consider the actual representation as used by the model for prediction. Especially considering the previous issue, we expect the model to have some kind of encoding to differentiate, whether the information of a dimension is present or not.

While we will never completely get rid of the first issue, we try to remedy misinterpretations coming from all three issues by closely analysing dimensions separatley in Section §4.2.3 together with the actual values of r in the dimension at hand. We reduce the shortcomings from the second issue by adding a filtering option, that we demonstarte in the next section by identifying dimensions containing semantic information.

Finding semtantic dimensions

Looking at the actual data we observe that several dimensions only include words referring to female humans. We investigate this finding by looking for dimensions that contain gender-specific information. Based on the data we create two wordlists for female³⁰ and male³¹ humans respectively. Following our observation, that more specific information causes more noise in the visualisation due to sentences not containing it, we only consider sentences containing at least one word from at least one wordlist. The resulting plot is depicted in Figure 9 with dimensions sorted by their informativeness according to SD. Words are labelled as *male* or *female*, if they occur in the according wordlist, any other word than those is labelled *OTHER*. Several interesting findings are shown here. The first two dimensions having the hihges variation do not distinguish between female or male humans, but instead jointly encode both information, seemingly focusing on humans having a gender. We will look closer into these dimensions in Section §4.2.3. More importantly however, we observe several dimensions with only male (red) and OTHER (yellow) words responsible for its value, while others only arise from female (green) and OTHER only. Comparing these dimensions we see, that some of them are strongly complementary to each other with respect to the encoded gender: We interpret dimensions that exclusively retrieve their values from female or OTHER words to encode whether there is a female human being present in the sentence (labelled as female) or not (labelled as OTHER). Similarly, dimensions exlusively arising from male or OTHER words are likely to encode whethere a male human being is within the sentence or not. Based on this interpretation one can observe that male-encoding dimensions are labelled as OTHER exactly in those dimensions, that are labelled as female in female-encoding dimensions and vice versa. Note that all displayed dimensions contain the highest overall SD, indicating that they are amongst the most expressive dimensions of the model. Being redundantly in several informative dimensions encoded, we conclude, that gender-specific information is highly relevant for SNLI.

³⁰ Words in the **female** wordlist: *daughter, daughters, female, females, girl, girls, lady, mother, mothers, her, herself, she, sister, sisters, wife, woman, women*

³¹ Words in the **male** wordlist: *boy, boys, dude, father, grandfather, guy, guys, he, him, himself, his, husband, male, males, man, men, son, sons*

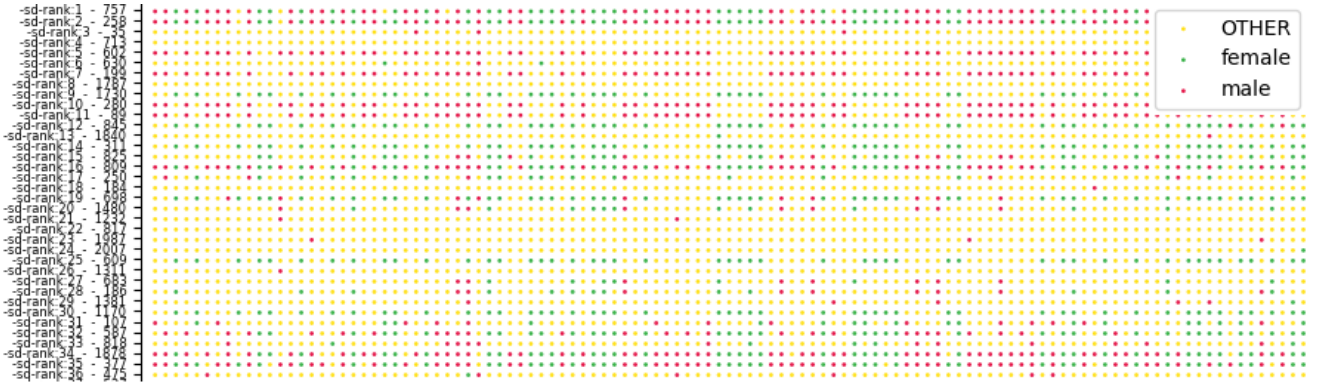


Figure 9: An extraction of a grid-plot, gender specific female using only sentences with words of pre-defined wordlists.

This observation is in line with Gururangan et al. (2018), who show that removing the gender information to create the hypothesis was a common heuristic applied by the annotators when creating the dataset.

4.2.3 Female and male dimensions

We rely on the method described above to manually find patterns that are encoded in the sentence representations and identify the corresponding dimensions. Following the drawbacks of our naive approach we conduct our dimension-wise analysis with respect to the actual values of the given dimension, that are retrieved from each word. We represent each dimension in a histogram by mapping the dimensional values from each sentence into a discrete space. In preliminary attempts tried to automatically divide the one-dimensional feature space of a dimension into meaningful intervals, however this did not create meaningful representations.

Female and male dimensions

Subsequently to the findings in the previous section, Figure 10 shows a detailed view of one of the male-encoding dimensions with no filter applied, thus using all 1000 sentences. For each sentence we retrieve the value of the two displayed dimension and assign it into bins of size 0.05 displayed on the x-axis. The amount of sentences containing a specific value is displayed on the y-axis for each bin, colored by the chosen labelling scheme. Based on our previous observation we assume the gender-dimension to only encode whether a human with the given gender is within the sentence or not. To ensure, terms for humans without a gender are not presented in those dimensions we create third category, containing words for humans regardless of their gender³². Additionally we move pronouns from our previous wordlists into new categories. Both dimensions undermine our initial assumption that they encode whether a male

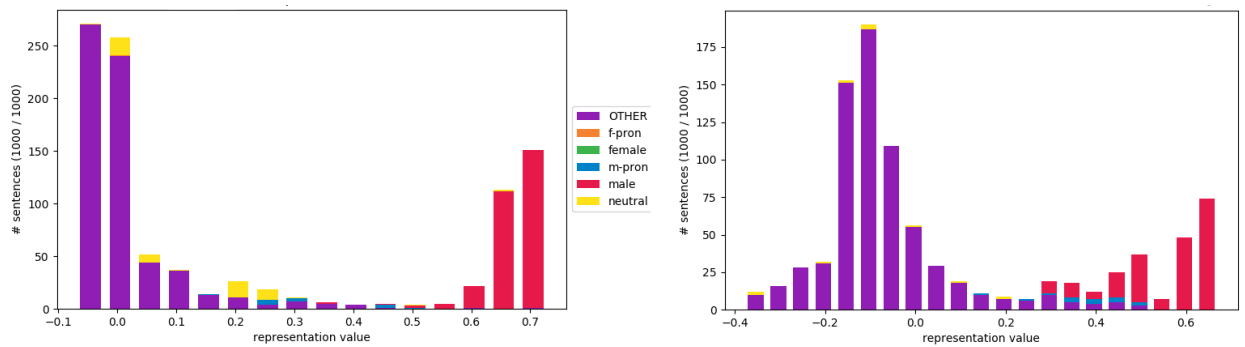


Figure 10: Representation visualization with respect to genders of dimension 199 (left) and dimension 602 (right).

human is present within the sentence or not. Clearly this is separated by the value within the dimension. While all high values arise from male words, most of them covered by our rather limited wordlists, not a single value stems from any word of the female wordlists. While neutral words may be responsible for values within these dimensions, their influence is negligible. All words coming from lower-valued bins seem arbitrary, arising from the fact that some sentences do not contain any male words and thus a random word will take its place, resulting in a low value. Even though the distributions

³² Words in the **gender-less** wordlist: *parent, parents, friend, friends, person, people, family, student, adult, adults, couple, couples, child, children*

are different, both presented dimensions seem to encode the same information. For an even more detailed view, we focus on the individual words from our male wordlist in Figure 11. We observe, that both dimensions enable a fine-grained

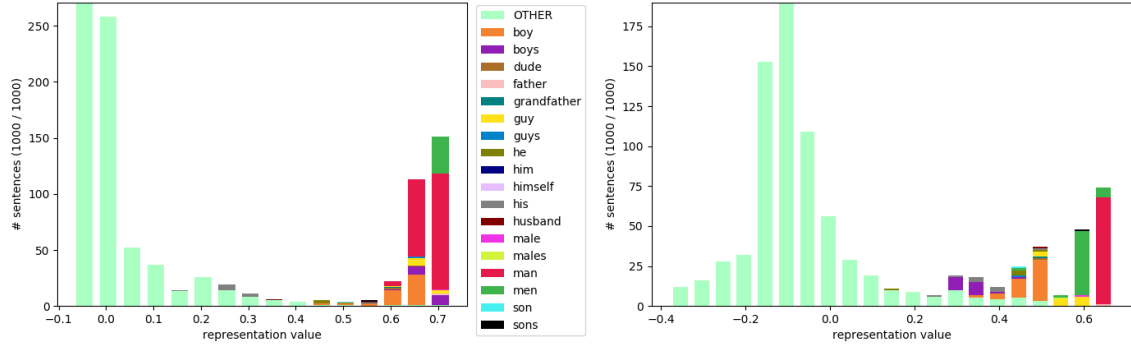


Figure 11: Detailed representation visualitation of different terms for human males of dimension 199 (left) and dimension 602 (right).

differentiation between different words and their meanings, even within their high values. In both cases, *boy* scores a lower value than *man* as being “less male”, indicating that this dimension not corresponds to the biological male gender but to attributes are generally assoziated with males. This only seems logical, as it is known that gender information is present in distributed word-representations (Mikolov et al., 2013b) that our model relies on. These again are determined by their surrounding context, which obviously is dominated by male-*assoziated* words. While both dimension seem to encode the same information, based on the words reaching high values, the encoding of this information slightly differs. Dimension 199 has the tendency to score higher values if multiple males, namely *boys* and *men*, are present, whereas dimension 602 reduces the value for plurals.

We similarrily investigate the female-endocding dimensions, depicted in Figure 12 already using the detailed labelling and observe the same principal encoding scheme. As for the male-encoding dimensions, all higher values within both

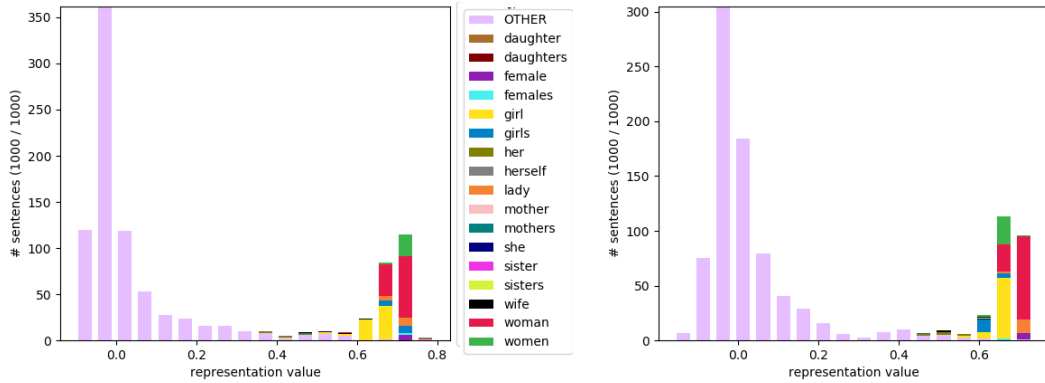


Figure 12: Detailed representation visualitation of different terms for human females of dimension 845 (left) and dimension 311 (right).

dimensions arise almost exclusively from our female wordlists. Younger females, namely *girl[s]*, are encoded using a lower value than *woman* or *women*. Furthermore, the different encoding of both female dimensions of singular and plural is aligned with the differences depicted in the male dimensions. Subsequent visualisations of other dimensions show similar results, such that highr valued words may easily grouped by some attributes while low values words seem rather arbitrary. We conclude that a dimension, encoding any kind of specific information *i*, by ssigning high values, if *i* is present within the sentence and low values if *i* is not given. Note that this explanation intuitively can be aligned with the max-pooling. Given that *i* is within the sentence, the model adjusts its weights such that the dimensions encoding *i* will result in high values, which naturally will be selected as being the highest amongst all values. In case *i* is not within the sentence any other arbitrary word will have the highest value of the specific dimensions, however this will be significantly lower than in the previous case. We show further evidence for this explanation in some remaining subsections.

Relevance of female and male dimensions

We want to gain information of how relevant those identified dimensions actually are when the predicting the relations on SNLI. Thus we conduct an experiment with sentence representations solely consisting of dimensions that we identified to encode gender-specific information. Specifically we found four dimensions to encode each gender respectively. We only consider subsets of these dimensions as sentence representations and train for each subset, consisting of an equal amount of female and male dimensions, a new model for 5 iterations using the same hyperparameters as in the Shortcut-Stacked Encoder[†]. Solely the size of the hidden layer is changed with respect to the size of the sentence representation due to being tremendously reduced. The results in Table 4.2.3 show that a sentence representation solely consisting of one dimension per gender reaches an improvement of about 9 points in accuracy over a random baseline with 33.33%. Adding more dimensions somehow reduced the additional improvements, indicating some redundancy but also some distinct information encoded within those dimensions. About half of the data can be classified correctly based on only 8 dimensions encoding gender-information of the sentence. This may indicate that these dimensions are highly relevant within the model, however all these observations may also arise from the model learning patterns, that are not considered by the original model. We thus conduct another experiment to shed more light into the impact of the detected dimensions.

$ r $	MLP size	Acc. (train)	Acc. (dev)
2	6	42.26%	42.87%
4	12	47.11%	47.41%
6	18	47.50%	48.64%
8	24	49.54%	50.37%

Table 5: Accuracies achieved on SNLI using $|r|$ -dimensional sentence representations of gender-specific dimensions.

Inverting gender information in the sentence representation

Knowing what information is encoded and how this is done we try to twist the sentence representations before they are fed into the classifying MLP, to see if it is possible to exploit the gained knowledge for adjusting the represented meaning without adapting the actual input sentences. Specifically we try to invert the meaning of the found gender-specific dimensions: If the sentence representation originally contains information that a male or female human is present, we change it to be not present and vice versa. Let d_j^i denote the value of the i th dimension within the j th sentence representation. For all i referring to gender-encoding dimensions³³ we calculate the maximum reached value d_{max}^i and minimum reached value d_{min}^i over all n sentences:

$$d_{max}^i = \arg \max_{v^i} (v^i | v^i \in \{d_0^i, d_1^i, \dots, d_{n-1}^i, d_n^i\}) \quad (6)$$

$$d_{min}^i = \arg \min_{v^i} (v^i | v^i \in \{d_0^i, d_1^i, \dots, d_{n-1}^i, d_n^i\}) \quad (7)$$

We further calculate the new value \bar{d}_j^i , replacing the original value d_j^i for all relevant i using the following equation. Note that we replace d_j^i by \bar{d}_j^i prior to the feature concatenation, ensuring to overwrite all relevant features. Basically we mirror each value on the dimension's mean, ensuring that the resulting values are within the common range for the given dimension.

$$\bar{d}_j^i = \frac{d_{max}^i + d_{min}^i}{2} + \left(\frac{d_{max}^i + d_{min}^i}{2} - d_j^i \right) = d_{max}^i + d_{min}^i - d_j^i \quad (8)$$

Rather than using the sample mean from all sentences, which would be heavily influenced by how much the encoded information is represented within the data, we calculate the mean based on the outer values, intending to focus on the information-encoding aspect. Considering the distributions of the dimensions having two peaks, either low or high valued, we assume this method to be appropriate for our experiment.

Evaluation of inverted gender-dimensions

We use the evaluated the proposed method in the full SNLI train and dev data and report our results in Table 4.2.3 together with the original performance of the used Shortcut-Stacked Encoder[†]. The table shows a range of experiments with an increasing number of dimensions being inverted on either both sentences or only the hypothesis. It can be seen that even after inverting all four dimensions for each gender, the performance only slightly drops, when applied on both sentences. This indicates that the performed equation (8) indeed is sufficient to invert the encoded information to a high

³³ **Male** dimension indices: 89, 199, 280, 602; **Female** dimension indices: 311, 609, 845, 1730

Inverted dimensions	Inverted sentences	Acc. (train)	Acc. (train) +/-	Acc. (dev)	Acc (dev) +/-
None	None	87.41	0.0	85.31	0.0
1 female, 1 male	premise, hypothesis	87.36	-0.05	85.25	-0.06
2 female, 2 male	premise, hypothesis	87.25	-0.16	85.19	-0.12
3 female, 3 male	premise, hypothesis	87.08	-0.33	84.97	-0.34
4 female, 4 male	premise, hypothesis	86.82	-0.59	84.78	-0.53
1 female, 1 male	hypothesis	87.05	-0.36	84.73	-0.58
2 female, 2 male	hypothesis	84.76	-2.65	82.29	-3.02
3 female, 3 male	hypothesis	80.23	-7.18	78.28	-7.03
4 female, 4 male	hypothesis	73.20	-14.21	71.69	-13.62

Table 6: Results in terms of accuracy of inverted gender-specific dimensions on SNLI train and dev set.

degree, since inverting the gender in p and h simultaneously should not have an impact on the final prediction³⁴. More importantly, when only the hypothesis’ representation is changed, we observe that inverting a single dimension reduces the model’s performance only slightly. An increasing amount of inverted dimensions also increases the impact on the overall accuracy. We conclude that this is due to redundant information, most likely coming from dropout.

Analysis of inverted gender-dimensions

In order to actually see that sentence meanings shifted according to our expectations, namely male individuals should be intereted as female and vice versa, we analyse the actual predictions of the data. the results for our observations when looking for differences between the models prediction using the untouched or inverted (using all eight dimensions, inverting the hypothesis only) sentence representations are shown in Table 4.2.3. The upper part of the table depicts samples with human main actors having gender that re correctly classified by the original model. By inverting the mentioned dimensions we invert the gender-aspect of these actors abd subsequently a *woman* in the hypothesis is actually interpreted similarly to a *man* by the model. We observe that the vast majority of samples containing the same gender in premise and hypothesis flip the predicted label after following this interpretation. Some samples without explicit gender information, like *dog* in thwe lower part of the table, remain with the same label. Especially for human

Premise	Hypothesis	Prediction (original)	Prediction (inverted)
A blurry <i>woman</i> eating fish.	The <i>woman</i> is eating dinner.	neutral	contradiction
A <i>woman</i> practicing for tennis.	A <i>woman</i> practices tennis	entailment	contradiction
Three <i>men</i> sitting behind a building.	Three <i>men</i> are sitting.	entailment	contradiction
A <i>male</i> in a green jacket points an imaginary shotgun at the sky.	A <i>woman</i> in a green jacket pointing an imaginary gun at the sky.	contradiction	entailment
A young <i>woman</i> with a ponytail climbs a white stone structure.	A young <i>man</i> has a ponytail.	contradiction	entailment
A little <i>girl</i> in brown is playing with two hula-hoops.	The person playing with hula-hoops is <i>male</i> .	contradiction	entailment
Two dogs standing in the snow.	The dogs are looking in the same direction.	neutral	neutral
Two people dancing outside.	Two people dancing.	entailment	contradiction
A country band is playing.	A group is playing music.	entailment	contradiction

Table 7: Comparison of samples between their predictions based on the original and gender-inverted sentence representations.

main actors without a specific gender, we observe unexpeted predictions. One possible explanation could be, that since words like *band* or *people* have no gender, the inversion of the hypothesis suggests that people of both gender are present in the sentence. Yet this would not explain that something similar does not happen in the *dog* sample.

While this experiment intentionally not improved the accuracy on SNLI, it shows that it is possible to adjust the sentence representation in a meaningful manner, using the gained insights on how information is encoded. The results give evidence that in the majority of cases, the new meaning corresponds to the initial intention, yet also comes with some minor side effects, undermining the need to have a deeper understanding how the representations are in fact used by the classifier.

³⁴ We select only 8 out of 2048 dimensions, that are highly representative for the gender-specific meaning. Correlated information however also has an impact on other dimensions that we left untouched, thus we don’t claim to have inverted the full gender-specific meaning, but a crucial amount of it.

4.2.4 Other semantic dimensions

Following the idea of high valued words being representative for the information encoded by a dimension we analyse more than 100 additional dimensions, finding mostly semantic relations between the words of interest. We give an overview about the semantic aspects covered in the representations and provide sample words taken from a single dimension each time.

- **Mixed:** The vast majority of dimensions encode even within higher dimensions several different ideas, that can be grasped by looking at the words. For instance one dimension simultaneously considers words as relevant that are related to fast movement or lonely emotion associations (running, runs, race, jogging, no, alone, dry, empty, crying, timid). Another dimension assign all very high values to nouns that are possible arguments to the word *play*, namely instruments and sports, somewhat lower but still very high words reflect associated verbs or tools for sporty or artistic activities (soccer, football, baseball, drums, tennis, accordion, guitar, saxophone, dancing, swimming, singing, painting, boat, bicycle, surfboard). In fact, most dimensions contain words within their high values that may easily be clustered in several groups. Sometimes however a dominant common meaning exists. All examples shown below of course include words that may be grouped as an additional category and not necessarily are directly related to the assumed encoded information. However if there is a highly dominant pattern, we ignore words that are divergent to this meaning, considering it as noise. Given the fact that the representation actually considers the context around the responsible word, we are only able to get an impression of the meaning rather than a accurate definition anyway, by solely looking at the individual words.
- **Community:** Several dimensions encode different communal aspects like family related topics (children sibling, mother, wife, school) or social events (friends, championship, lunch, family, baseball, party)
- **Children:** Different dimension represent children in varying contexts. These dimensions show, that indeed context is captured within the dimensions and they not solely rely on the word we investigate, yet they can be interpreted. Specifically we find several dimensions with children in playful contexts (boy, girl, young, playing, game, teens, skateboarding, soccer) or in the context of caring and comfortness (boy, girl, young, child, sleeping, hungry, tiny, napping, sad, asleep)
- **Locations:** A huge amount of dimension encodes locational information. This may be for instance city or building like (pool, inside, restaurant, sidewalk, floor, bar, classroom, museum, downtown, building) water oriented (beach, pool, water, lake, river, mud, ocean) or referring to different grounds (street, beach, road, sidewalk, grass). Several dimensions however show topic related locations together with possible activities and are harder to categorize (street, beach, outside, park, soccer, rock, truck, boat).
- **General atmosphere** Some dimensions consist of a broad range of words, however it still is obvious that there is a higher common meaning. One dimension for instance ranges from activities to locations or food, all seemingly indicate some level of lazy comfortableness (sitting, inside, sleeping, bed, room, dinner, cream, milkshakes, doll).
- **Activities** Several dimensions include some kind of activity related words, consisting of both, verbs and nouns, clearly showing common attributes (ball, game, race, competition, skateboard, concert, artist), or encoding verbs that usually take positional arguments (walking, sitting, running, standing, walks, jumping), while others solely focus on only one of these meanings (standing, stand, stands, feet).
- **Others:** We found more dimensions not fitting into a broader category but still encoding a very specific information. For instance one dimension considers everything that has to do with “wearing clothes” as a high value (wearing, dressed, covered, shirt, umbrellas, naked, jersey, dress). Another dimension clearly consists of terms describing humans for a profession or activity (player, vendor, skier, musicians, clown, workers, jockeys, artist).

While it usually is hard to specifically name the attribute that most of the high valued words within a dimension have in common, it is usually very straightforward to grasp the general idea. Yet these words give valuable information that enables us to interpret sentence representations, a large advantage considering that initially nothing was known. Almost exclusively we found these common ideas to be based on the semantics.

4.2.5 Syntactic dimensions

After extensively looking for semantic information in the sentence representation, we investigate how much syntax is encoded by looking for POS and dependency parse labels.

Verbs and adjectives

We start by syntactic patterns across the sentences looking for verbs and adjectives using POS tags. One finding that is highly dominated by verbs is depicted in Figure 13 (left). Looking at the actual words that are responsible for high

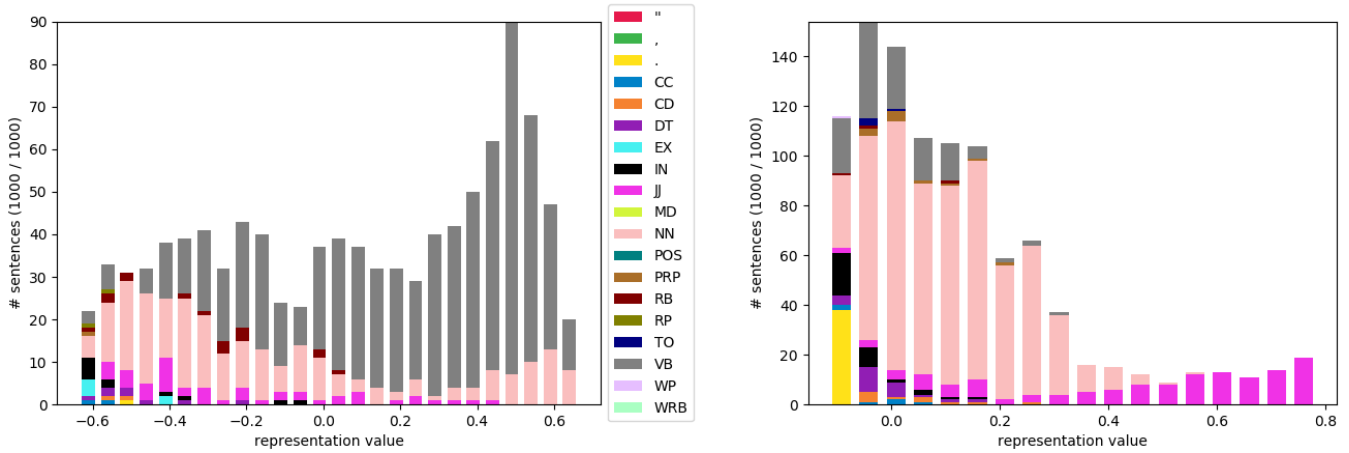


Figure 13: Dimension 713 encoding verbs (left) and dimension 2020 encoding adjectives.

values within the dimension however, we find that there is also a semantic commonality between them with verbs mostly being *playing*, *walking*, *sitting*, *running*, *standing*, *swimming*. Several interpretations on what they have in common are plausible, like all taking locations or places as arguments or all being some kind of physical activity³⁵. Nouns scoring high values exclusively denote sport types like *football*, *basketball*, *tennis*, *baseball*, *volleyball*, *hockey*, which represent in combination with a verb also a physical activity. Looking at the words of the adjective-encoding dimension, depicted in Figure 13 (right) shows an even more obvious semantic relation, since all adjectives³⁶ are typically used to describe people, even though many different adjectives do exist in SNLI. We face the same problem as described earlier, that different attributes correlate, making a definite interpretation impossible. We also observe that this dimension, as well as others, differs strongly with the gender-specific dimensions in their distributions. Gender-specific dimensions consist of two peaks, intuitively because they encode basically two states, either the gender is present or not. This dimension however is encoded using a relatively large range of values, all being relatively equally represented. Even though we do not go deeper into a fine-grained analysis of dimensional values, due to the impact arising from encoded contexts, we assume that other information, as in this case, can be scaled.

Subjects and objects

The differentiation between subject and object that can be identified using dependency parsing seems highly useful for classifying image captions. While the subject most likely refers to the main object, depicted in an image, the object may serve as a more informative explanation, however most likely being less relevant. To identify whether the model learns equivalent information, we look for subjects and objects respectively. We also look for predicates, however this is especially noisy since many sentences in SNLI actually are noun phrases and thus lack a verb. The closest to encoding this information, even by using dependency parsing labels, was the dimension in Figure 13. We find the two dimensions in Figure 14 for both, subject and object, respectively. Similarly as when analysing the verbal dimension, the first sight suggests that indeed information as from a parse tree is encoded within those dimensions. While this actually may be true, undeniably both words retrieve high values from words that are also semantically highly related. High dimension of the dimension seemingly encoding subjects (left) exclusively contains words referring to people³⁷ when having high values. While they all are very likely to encode a very important aspect and subject of the sentence, subjects referring to anything else than humans are not considered by this dimension and it seems more likely that the semantic relationship is encoded and just happens to be the subject. The object-encoding dimension (right) shows the same phenomenon, solely encoding words referring to places³⁸ with a high value.

We clearly see that syntactic information is indeed encoded, both for the dependency parse information as well as POS. Yet those dimensions highly correlate with semantic and it seems much more likely that the identification of these

³⁵ Having different scales how physically intensive it is, in the sense of being sporty.

³⁶ High valued words of the **adjective** dimension: young, little, old, small, older, fat, large, elderly, lean, middle-aged, ...

³⁷ High valued words of the **subject** dimension: man, woman, girl, boy, men, women, boys, girls

³⁸ High valued words of the **object** dimension: street, beach, pool, outside, park, road, restaurant, sidewalk, grass, city, ...

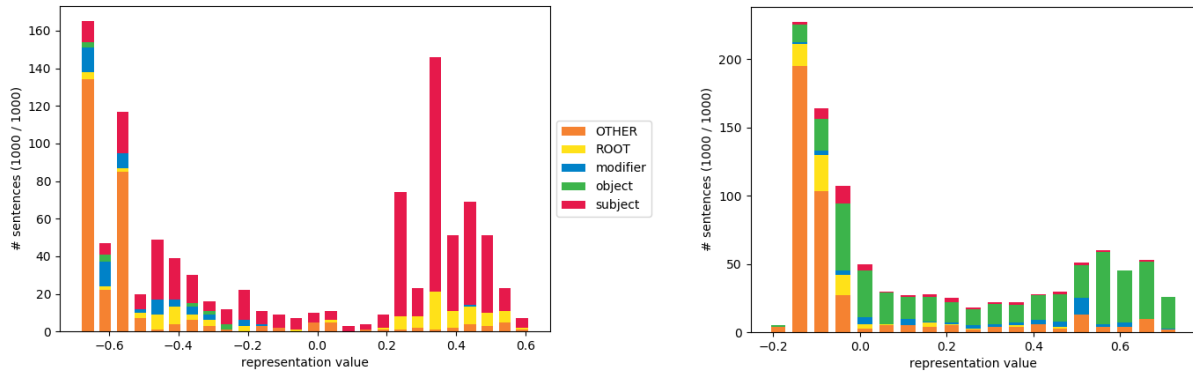


Figure 14: Dimension 757, encoding the subjects (left), and dimension 1840 encoding objects (right) of sentences.

semantic patterns is sufficient for the model to rely on it to encode syntactic information. This obviously is coming from SNLI with the majority of sentences regarding people. Another plausible interpretation is that the model does not actually require any syntactical knowledge for the task, based on the simple sentence structure. We conclude that whether the model indirectly uses syntactic information, originating from semantic features, or solely leverages semantic information, not relying on syntax at all, is a matter of perspective; the truth lies probably somewhere in between.

4.3 Insights on the sentence alignment

We have shown that dimensions represent a specific information of any kind. High values within these dimensions indicate that the information, encoded by the given dimension, is present while low values indicate it is not present in the sentence. In this section we analyse, using the newly gained insights, how the models finally align the encoded information in the sentence representations to predict the entailment relation label. For our analysis we sample 150 premises, each with one hypothesis for each label respectively, that are all classified correctly by Shortcut-Stacked Encoder[†].

4.3.1 Alignment analysis on a single sample

We first analyse single samples in order to identify plausible strategies for the network, when mapping both sentence representations, knowing only very little how the network actually leverages from the information in the premise. We demonstrate our results with the samples premise:

Premise: A woman sitting in the dirt.

and its three hypothesis, one for each label:

Entailment: There is a woman sitting *outside*.

Neutral: A *dirty* woman sitting in the dirt.

Contradiction: A woman *standing* in the sand.

We highlight the words within each hypothesis, that we consider relevant for the label, based on our human judgement. The entailing hypothesis describes for the most part the same setting as the premise. Since *dirt* usually appears *outside*³⁹, we consider this change of words as a generalization, meaning *outside* includes *dirt* (amongst others). The neutral hypothesis introduces new information about the woman being *dirty*, which is very plausible, yet not explicitly given in the premise. The contradicting hypothesis clearly is incompatible with the premise, as the woman can either be *standing* or *sitting*, but not both. In this subsection we show that indeed, the identified differences can be easily observed when looking at both sentences simultaneously for the entailing and contradicting hypothesis. For the sake of brevity we omit the analysis of the neutral sample in this subsection, as it does not give any additional insights.

Visualizing the entailment relation

We start by visualizing the relation between the premise and the entailing hypothesis. We assume that the model will rely on the information encoded within a given dimension to compare the meanings of two sentences and infer its relation. Figure 15 (left) shows the alignment between the premise (y-axis) and the hypothesis (x-axis) by

³⁹ Based on its context in conjunction with “sitting”, dirt is most likely used in the sense of being a dirty outside ground.

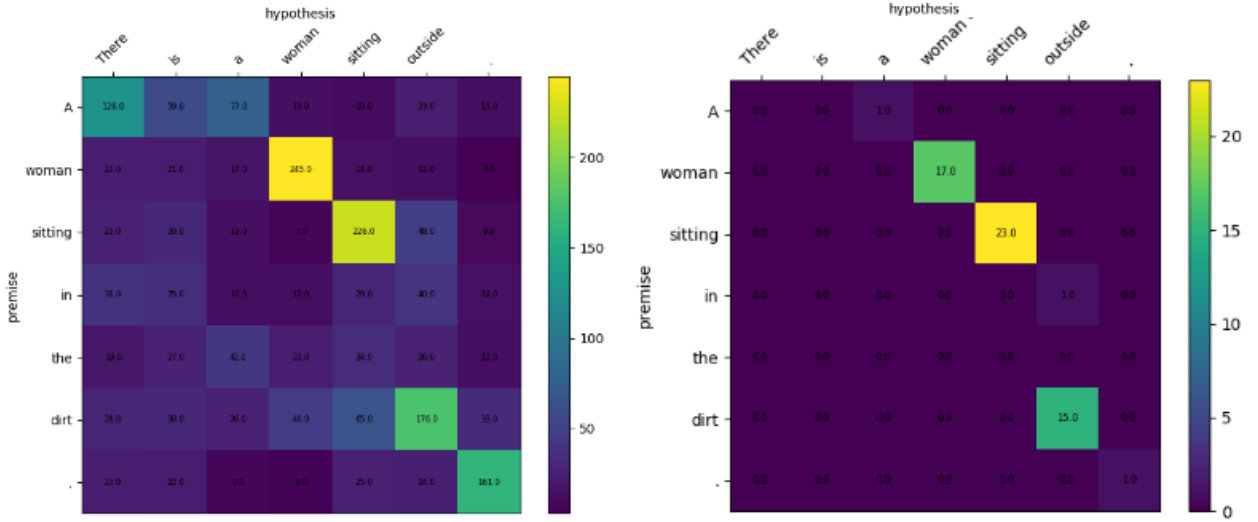


Figure 15: Word alignments of an entailing sentence pair either by counting all shared dimensions (left) or only dimensions with at least a value of 0.2 (right).

counting all dimensions d_i arising from each word. For each word we identify all dimensions d_i that are represented by it in the representation. The intersection of a word from the premise and a word of the hypothesis is the total amount of all dimension with the same identifier i , that both words represent. thus for instance *dirt* and *outside* have 176 dimensions in common. This plot is quite noisy, since it does not differentiate between different values, thus even dimensions that encode information which is not given in both sentences are arbitrarily aligned between two words. Following our insights from the previous section we filter out all dimensions that do not at least have a value of 0.2 in both sentence representations in Figure 15 (right), presumably resulting in only compared information that is present in both sentences. It can be observed that by applying this filtering the main aspects of both sentences ("woman/woman", "sitting/sitting", "dirt/outside") are strongly aligned with each other, while the remaining words only show little similarities w.r.t. their encoding. this indicates that aligning both sentences intuitively can result in the entailment label within the examined example.

Note that the actual Shortcut-Stacked Encoder does not only rely on the indial sentence representations but also combines them using element-wise multiplication and difference. While element-wise difference intuitively serves a direct

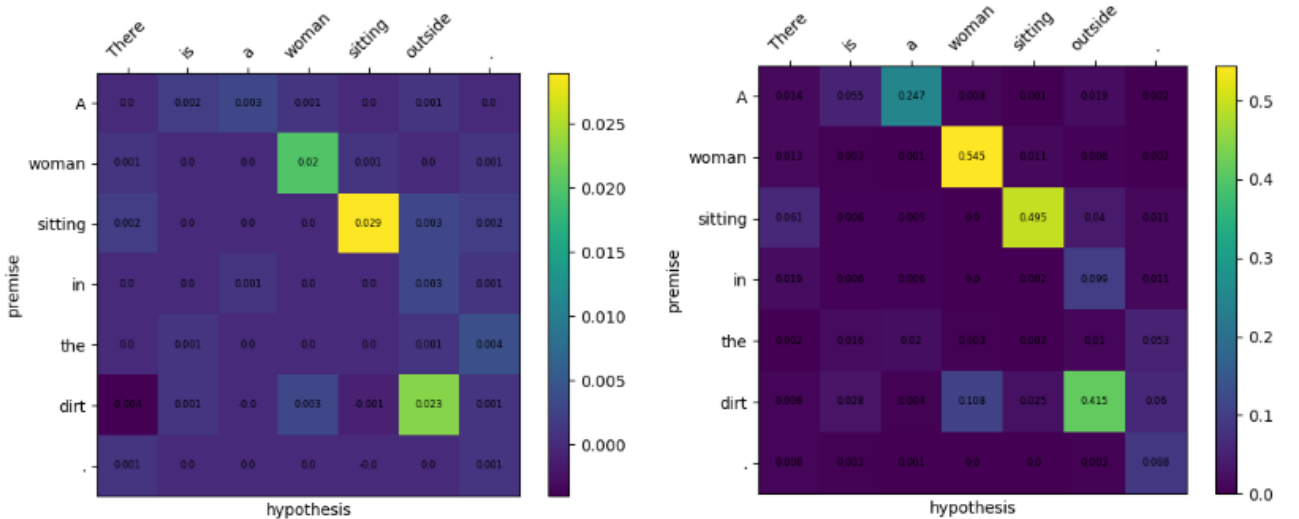


Figure 16: Visualisation of an entailing sample with applied element-wise multiplication either using the mean (left) or maximum (right) product of all shared dimensions for each word pair.

comparison of the encoded information per dimension, the effect of the multiplication feature is less obvious. Figure 16 (left) visualizes the mean product of both sentence representations using element-wise multiplication as described in

Section §2.3, averaged by the amount of shared dimensions as counted in Figure 15 (left). One can observe that the plot arising from element-wise multiplications similarly highlights the same relevant word relations as seen in the previous plot, indicating that it serves as some kind of soft AND-operator, as it is showing that similar information is present in both sentences. As this plot again might be heavily influenced by irrelevant relations we also visualize the highest elementwise product of all shared dimensions between two words in Figure 16 (right), showing a similar pattern.

Visualizing the contradicting relation

We visualize the shared dimensions of both contradicting sentences in Figure 17 (left) in the same manner as done for the entailment sample, by only counting dimensions that achieve a value of at least 0.2 for both words of each word-pair. Note that not only the entailing word-pairs show similar values within several dimensions but also *sitting* and *standing*

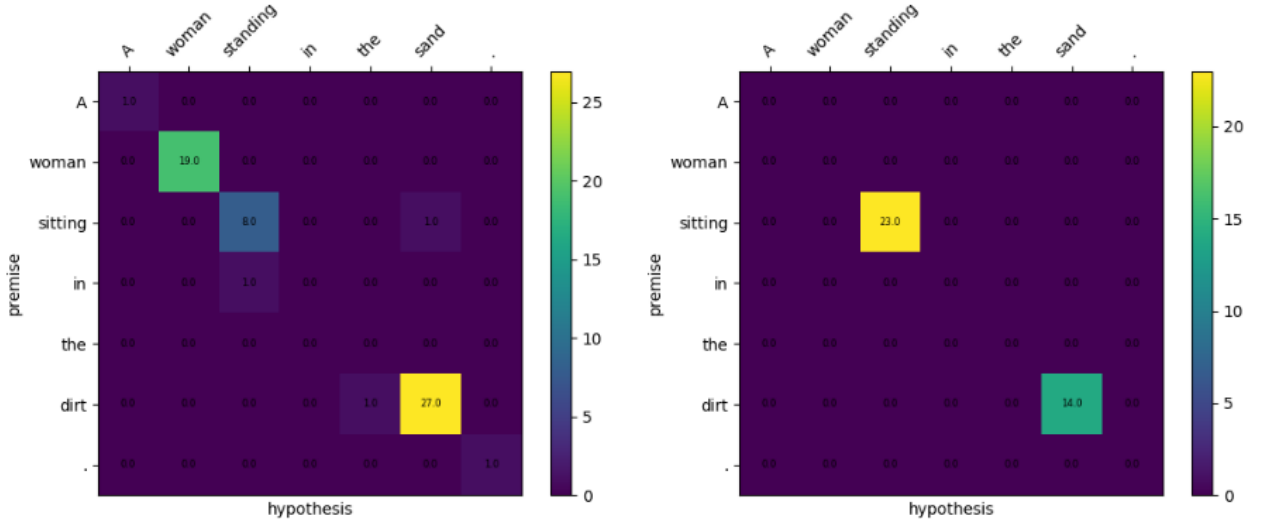


Figure 17: Visualisation of a contradicting sample by counting meaningful shared dimensions (left) and meaningful distinct dimensions (right) amongst pairs of words.

share the same meaning in some cases. This obviously makes sense, as both verbs are similar w.r.t. several aspects. Since the Shortcut-Stacked Encoder creates sentence representations without looking at the other sentence (without inter-sentence attention) it must encode all information that might be relevant and is not able to focus on specific relations that would be crucial for this particular sentence pair. Thus, we also count dimensions that are distinct between two word pairs, depicted in Figure 17 (right). This is the case for the majority of cases, especially since completely unrelated words are encoded by a high amount of distinct dimensions. In order to remove noise coming from this issue we apply two thresholds in our visualization:

- **Ensure meaningful relations:** To exclude the counts of dimension between unrelated words, we only consider word-pairs with at least 5 meaningful (in the sense of both values reaching at least 0.2) dimensions. This is motivated by the assumption that the model needs to learn which dimensions can be aligned in a meaningful way, which only is plausible if both words also share at least some commonalities.
- **Ensure meaningful value:** Only considering word-pairs with meaningful relations, we only count dimensions that are distinct for both words if the word of interest reaches at least a value of 0.2, thus encodes information.

We observe that indeed *sitting* and *standing* are encoded using a lot more different than shared dimensions and take a closer look at the actual dimensions in Figure 18. The plot shows all dimensions “meaningful” dimensions that only encode information coming from one of both words. Each dimensions show two bars, the left bar indicates the value within the representation of p , the right bar of h . Colors show the word, which is responsible for the value. Additionally we leverage information gained for some dimensions for the previous section, by providing sample words of the general meaning encoded by each dimension. We only show these indicators for dimensions that we labelled prior to the visualization to not be biased towards a specific interpretation. We observe that the identified dimensions indeed highly differ in their value and may be important features to detect contradiction. This plot also is in line with previous conclusions that not-given information results in low values coming from arbitrary words. We also see that the information of what is encoded in the dimensions is helpful for understanding the shown values and do correspond to the responsible word if having a high value.

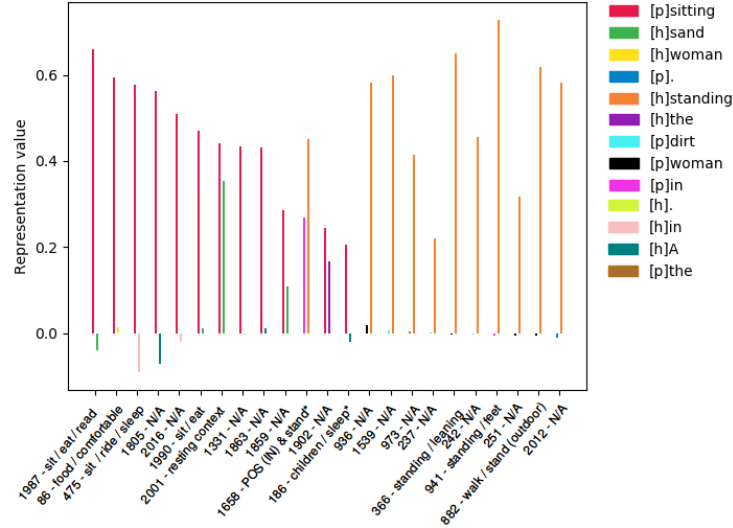


Figure 18: Dimension-wise visualisation of distinct information represented by *sitting* in the premise and *standing* in the hypothesis.

4.3.2 Approach for a general alignment understanding

Our previous results show that it theoretically is possible for the model to align relevant dimensions to infer the entailment label and suggest that doing so it can differentiate between contradicting or entailing meanings of two sentences. Other than the fact that the model predicted the examined sentence correctly however, our findings are more theoretical than how the model actually leverages this information. To also take into account the actual prediction based on the MLP we conduct another experiment. Based on our conclusions that high values indicate the presence of a specific information we formulate a very simple form of lexical entailment w.r.t. to our identified encoding schemes. Let I_p and I_h be the sets of information of any kind within p and h respectively. We further assume that lower values within a dimension generally represent less information w.r.t. the information encoded by the dimension:

1. The hypothesis contains a subset of information ($I_h \subseteq I_p$) would either result in paraphrasing ($I_h \equiv I_p$) or in less specific information in I_h , consequently being more general. We expect both cases to be labelled as entailment. We assume that for instance a hyponym *monkey* of the hypernym *animal* contains the same high dimensions as its hypernym and additionally more information that is specific for being a *monkey*.
2. The hypothesis contains a true superset of information ($I_p \subset I_h$) results in the opposite case of the one above. The additional information $I_h \setminus I_p$ is possible true based on the premise, yet not given. We expect this case to be labelled as neutral.
3. The hypothesis and premise contain different information ($I_p \not\subseteq I_h \wedge I_h \not\subseteq I_p$). We expect the model to predict contradiction if the amount of exclusive information in I_p and I_h is relatively large.

In subsequent sections we will refer to those assumption using their numbering (1), (2) or (3) respectively. First however, we demonstrate this intuition based on an artificially created sample:

Premise: A green man is running on the street.
Hypothesis: A man is running on the street.

This is predicted as entailment, as expected by assumption (1) by our model. After swapping the premise with the hypothesis, the predicted label is neutral, as expected by assumption (2). We visualize both sentences in Figure 19 and include additional hints to explain how this visualization can be read, as we will use the same technique when looking at multiple samples simultaneously. Intending the figure to serve the validation of our claims, we create it in the following manner: Let $D = \{i \in \mathbb{N} | 1 \leq i \leq n\}$ denote the set of all n dimensions. Furthermore let p_i and h_i denote the value within the i th dimension within p and h respectively. We divide the value range of all values in $\{p_i | i \in D\}$ and $\{h_i | i \in D\}$ respectively into a discrete space using bins of size 0.1, displayed at the y-axis for p and the x-axis for h with their lower bounds. For each $i \in D$ we identify the corresponding bins based on the values p_i and h_i and increment the intersecting field by one. Thus, for instance, 1171 dimensions have a value $-0.0241 \leq p_i < 0.0759$ for the premise, while also having a value $-0.0244 \leq h_i < 0.0756$. Following our insights, that low valued dimension encode the absence of information, we consider the upper left corner as irrelevant for the relation classification of both sentences. Arising from

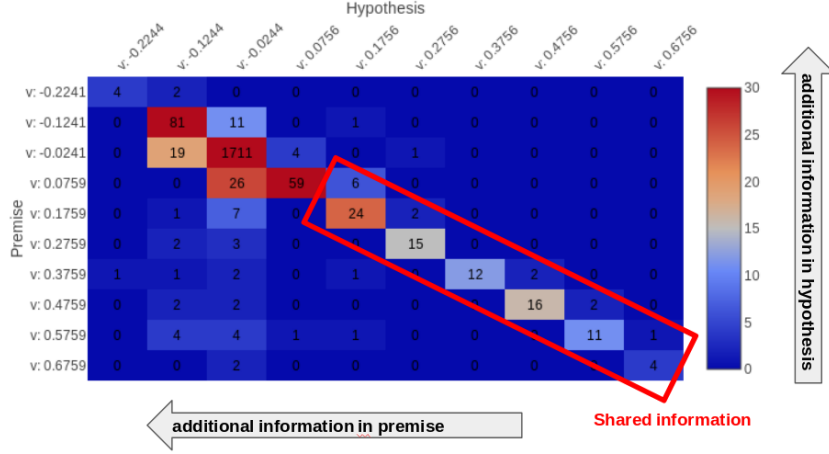


Figure 19: Visualisation of a sample sentence pair with explanatory guides for interpretation.

the same observations the diagonale, marked by the red recangle, corresponds to information that is present in both, p and h . Subsequently, everything that is above this diagonale represents information that only is present within h and accordingly, everything to the left of the diagonale is only present within p . We check the origin of all p_i to the left of the diagonale and observe that they exclusively emerge from the word *green*, which is the only additional information, given in p . Knowing that this naive assumption does not completely hold, we evaluate it on the chosen 450 correctly classified examples. While we find evidence for (1) and (2) in some cases, we will show why (3) is not sufficient.

4.3.3 Entailment analysis

We conduct the same experiment as conducted using a single sample in Section §4.3.2 over 150 correctly classified mples with the gold label *entailment*. The resulting plot in Figure 20 is calcualted identically as the plot with thee sample sentences, however displaying the mean amount over all sentence pairs rather than the the absolute amount. We observe that indeed, the majority of sentence pairs contains more information within p than in h , undermining our assumption (1). On the other hand only very few information are present in h and not in p . Yet, in order to get a better understanding and seperate entailment relations based on paraphrasing from entailment based on generalitzation we repredict the sentence pairs (p, h) with premise and hypothesis swapped as (h, p) . We expect all sentence pairs that are

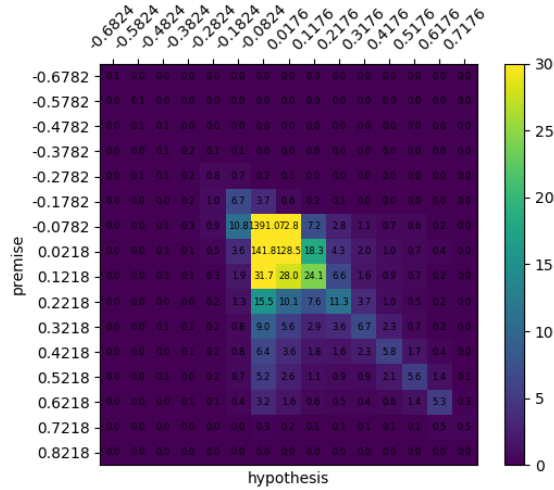


Figure 20: Visualisation of 150 sentence pairs (p, h) , correctly labelled as entailment.

predicted *entailment* for (p, h) and (h, p) to be paraphrasing. We consider all entailing sentence pairs that are *neutral* after swapping to arise from generalitation with the more general sentence h now being the premise. The resulting label distribution after swapping, based on the prediction of the Shortcut-Stacked Encoder[†] is listed below:

- **Entailment:** 11 samples (7.3%)
- **Neutral:** 111 samples (74.0%)

- **Contradiction:** 28 samples (17.7%)

Based on the model's prediction the majority of cases are described by the second scenario. Only very few show the same encoding in both sentences. To understand why some samples are predicted as being contradicting after being swapped we take a look at the actual data. As it turns out, in addition to misclassifications, many of the samples contain quite specific p with a highly general h , for example (before swapping):

Premise: A girl reaching down into the water while standing at the edge of a river.
Hypothesis: The girl is outside.

This should definitely fall into the case of our assumption (2), yet labelling it as contradiction if swapped to (h,p) may be correct, considering that while the more specific sentence p entails quite a lot information w.r.t. h , making it possible, yet not very likely. Considering that this is a plausible labelling, we conclude the first missing scenario of our assumption, a high amount of extra information in the hypothesis compared to a very broad claim in the premise may be classified as contradiction, rather than our assumption (2) neutral. Looking at the actual data that is now classified as entailment or neutral seems in line with our assumptions (1), (2). Since we are only focusing on these two assumption in this section, we visualize in Figure 21 samples predicted as entailment (left) and neutral (right) after swapping, since they describe the scenarios we initially intended. Indeed, samples that seemingly are paraphrased, based on the model's prediction, show

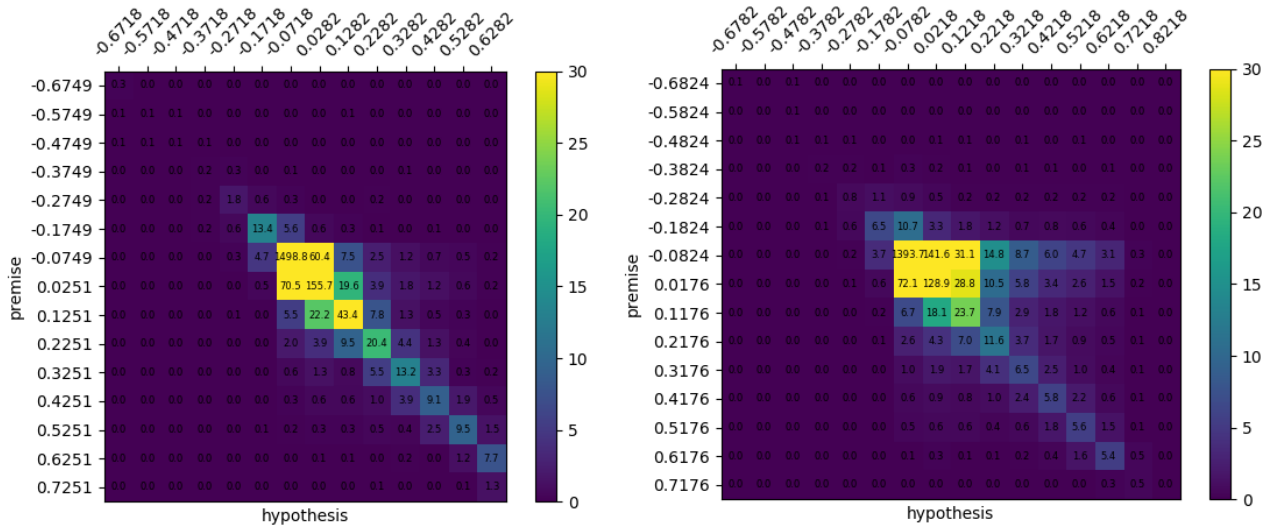


Figure 21: Visualization of samples predicted as entailment (left) and neutral (right) after swapping p and h .

highly identical meaning representations over all dimensions. Similarly, all samples that are now labelled as neutral and before swapping were considered entailment, the vast majority of cases, show a high tendency of encoding only a subset of information in of the new hypothesis within the new premise. Both plots show that both our assumptions (1) and (2) are correct, accepting the fact that other scenarios exist, as shown by some swapped contradicting samples.

4.3.4 Neutral and contradiction analysis

Our previous analysis showed that we can indeed observe how the model identifies the entailing label and to some extent how this differs from neutral and, in one specific setting, from contradiction. We now try to get more insights on neutral and how it differs from contradiction based on the sentence representations. The results (without swapping) for 150 correctly classified neutral and contradicting examples respectively are displayed in Figure 22. Unfortunately, we do not likewise find any patterns and also do not succeed by separating them for a more detailed analysis into smaller subgroups. This does not mean that our assumption (3) is incorrect, since indeed contradicting samples show very distinct information in p and h . Yet the same thing obviously happens for the neutral label. It might be slightly less distinct information when looking at the absolute numbers, however this difference is far from being representative. We find an explanation for this issue by looking into the samples. Consider the following two samples, classified correctly as neutral:

Premise	A group of kite surfers are busy surfing some waves.
Hypothesis	The kite surfers are participating in a race.
Premise	A baby laughing on the floor.
Hypothesis	A baby is being tickled.

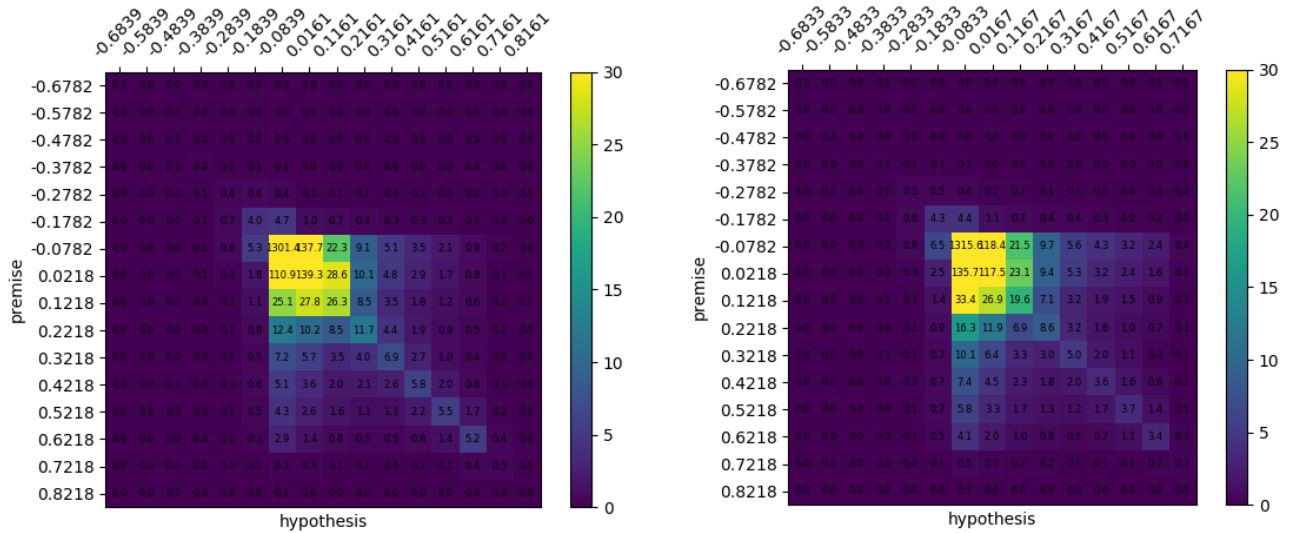


Figure 22: Visualitazion of 150 sentence pairs (p, h) correctly labelled as *neutral* (left) and *contradiction* (right).

In both cases, premise and hypothesis do contain distinct information, yet this information is not mutually exclusive but highly compatible with each other and in fact may also be true. We also look at correctly classified contradicting samples:

Premise	A boy eating at a table.
Hypothesis	A boy coloring at a table.
Premise	A baby laughing on the floor.
Hypothesis	A toddler is crying.

Also in this case both sentences encode different information, this time however it is not compatible, as the baby is either *laughing* or *crying* and the boy is either *eating* or *coloring*, but not both things. We see that assumption (3) holds for contradiction in terms of recall, however it is not sufficient to distinguish between neutral and entailment. We conclude that in order to do so, it has to be identified which dimensions can be true simultaneously, indicating neutral, and which dimensions exclude each other, indicating contradiction. in order to see how this is done within the model we need to understand the MLP on top of the sentence representation. This however goes back to the main drawback of standard neural networks being hard to interpret and we end our model understanding analysis at this point.

4.3.5 Summarizing the insights on max-pooled sentence representations

We have shown that it is possible to identify the general meaning of a dimension based on the words of a sentence, responsible for the dimensions value. We observe that high values within dimensions encode the presence of information, which intuitively goes with the nature of the max-pooling mechanism, while low values indicate the absence of information. In an experiment we additionally have shown that knowing the encoding-scheme and information of a dimension, it is possible to change sentence representations in meaningful intended way, yet observed minor side-effects. in the second section we showed that by aligning dimensions and knowing their encoded information, it is possible to understand sentence representations well enough to interpret the prediction reason. In the last part we analysed how the model actually performs the alignment between the sentence representation and showed intuitive explanations for some cases of relatedness between both sentences.

All these results are based on a rather small amount of rather small sentences, stemming from SNLI and thus merely show experimental results. For a general claim similar experiments should be conducted on larger size and range of data, especially containing longer text, using different models (all with max-pooled sentence representation) to see if those claims still holds.

4.4 Identification of missing knowledge

In order to see what information is not captured by the model and can be helpful for integration we analyse the errors made on SNLI. This analyses uses Shortcut-Stacked Encoder^{††}, being closer to the reported accuracy by Nie and Bansal (2017).

4.4.1 Approach

Aiming for missing knowledge and not for a general error analysis we focus on misclassified samples with gold label contradiction, predicted as entailment and vice versa, as we find that the label neutral was overall not that well understood by the SNLI annotators. To simplify the process we sample according to this constraint sentence-pairs having a high lexical overlap⁴⁰ and look for the knowledge required by a human to predict the correct label. In total we look at 196 samples incorrectly predicted as entailment and 163 samples incorrectly predicted as entailment identify common categories of missing knowledge. Note that by prefiltering the samples this way, our results exclude certain aspects. For instance many contradicting samples have multiple exclusive elements, resulting in a low lexical overlap.

4.4.2 Quantitative results

We show our quantitative results with the identified categories including sample sentence pairs for both labels in this section.

Misclassified contradicting samples

Table 4.4.2 shows the misclassifications of samples labelled incorrectly as entailment. We find that most problems arise

Problem	Type	Amount	Example
cohyponym	Nouns	29	A <i>creek</i> runs through the grassy area. A <i>lake</i> appears in the grassy area.
cohyponym	Verbs	32	The boy is <i>riding</i> his skateboard. The boy is <i>carrying</i> his skateboard with him.
cohyponym	Amounts	33	<i>Three</i> people resting on a snowy mountain. <i>Four</i> people are on a snowy mountain.
antonym	Adjectives	13	The ground is <i>covered</i> in snow. The ground is <i>visible</i> .
antonym	Verbs	14	boy <i>pushing</i> wagon with two pumpkins in it A boy is <i>pulling</i> a wagon with two pumpkins in it.
antonym	Prepositions	17	A man walking <i>down</i> stairs. The man is walking <i>up</i> the stairs.
structure	All	15	a sheep chases a dog. There is a dog chasing a sheep.
common sense	All	20	Someone in a 3ft swimming pool. A person is in a very large and deep pool.
<i>ignored</i>	<i>ignored</i>	23	A man climbing a rock wall. A man climbs the wall.
Total	-	196	-

Table 8: Misclassified samples with gold label *contradiction*, predicted as *entailment*.

from words sharing lexical relations, namely antonymy or cohyponymy for different kind of POS. We opt to show *Amounts* as a separate category, due to its high frequency. While Verb antonyms may also be considered as cohyponyms, we also list them separately, if they refer to the opposite meaning. Some samples, listed under *structure* require the model to take word order into consideration, and is mostly represented by semantic role reversal. We assign all samples to *common sense*, that requires information that cannot be retrieved using lexical semantic relations and usually need additional information, implied by the described entity or activity. Any sentence pair were we could not identify the required knowledge, due to not agreeing with the label, highly ungrammatical sentences or some rare very specific details are ignored in our results, categorized as *ignored*.

Misclassified entailing samples

Analysing the required knowledge for misclassified entailing samples as contradiction, we identify different categories, depicted in Table 4.4.2. Essentially we find three different ways, of how the relation between p and h can be described

⁴⁰ Only considering a sentence pair if at least 50% of the words within the shorter sentence are contained in the other sentence in a BoW perspective. Casing is ignored.

Problem	Required knowledge	Amount	Example
Paraphrasing	lexical knowledge	16	Two people play <i>foosball</i> . Two people are playing <i>table soccer</i> .
Paraphrasing (negated opposite)	lexical knowledge	19	A young boy is <i>sleeping</i> . A child is <i>not awake</i> . girl <i>opening</i> cosmetics <i>bottle</i>
Paraphrasing	world knowledge	20	The girl is <i>removing the top</i> off the <i>bottle</i> .
Generalization	lexical knowledge	30	The two <i>boxers</i> are females. There are two female <i>athletes</i> .
Implication	world knowledge	53	A hockey player makes a shot. A hockey player <i>is on ice</i> .
<i>ignored</i>	<i>ignored</i>	25	two people sit on a bench. two people sit on sand near water.
Total	-	163	-

Table 9: Misclassified samples with gold label *entailment*, predicted as *contradiction*.

and we differentiate between lexical- and world-knowledge being required for the correct prediction. In the example of paraphrasing, for *foosball* and *table soccer* or *sleeping* and *awake* it is sufficient to detect that they are synonyms or antonyms respectively. The third case required some actual understanding of the process of *opening* in the context with a *bottle*. While some of these samples may be understood by knowing several meronyms, we consider them to require a deeper conceptual understanding of how things work. The largest group also is interesting, as all *h* contain additional information, not directly given by *p*. While this case should usually be labelled as neutral, in this case the additional information is automatically implied⁴¹ (even though not textual) by *p*. This implication also required other external knowledge than lexical relations.

4.4.3 Conclusions

We observe that especially the classification of contradicting samples can be improved by using lexical relations, which are available in WordNet, as described in Section §3.1.1. While those may partially be helpful for the entailing samples, in their case world-knowledge seems more relevant, as could be gained from resources like Wikipedia (see Section §3.1.2). While obviously the goal of NLI is, to have proper reasoning capabilities, capable of dealing with world knowledge, we emphasize the aspect of lexical knowledge, arguing that any insights on how to incorporate this (simpler) information may later be applied used to include world-knowledge.

Note that the identified problems with lexical relations, especially *antonymy* and *cohyponymy* in Table 4.4.2, refer to the same problem, stemming from the nature of distributed representations. Essentially, they follow the distributional hypothesis, described by Pantel (2005) as

“[...] words that occur in the same contexts tend to have similar meaning.”(Pantel, 2005)

Considering a sentence like “The president of Italy hopes to get re-elected.”, one can easily conclude that *Italy* is a country based on its context. Even replacing *Italy* by any made-up country would intuitively result in the same conclusion. Subsequently to being represented by their context words, distributed embeddings supposedly have better generalization abilities (LeCun et al., 2015) due to encoding this phenomenon. However also mutually exclusive co-hyponyms or antonyms often share similar contexts (like most countries could replace *Italy* in the given example), resulting in very similar vector representation despite opposite meanings. This is a known problem (Sahlgren, 2008) of distributed word representations and several approaches, as explained in Section 3.4.1, aim to fix these problems in the embedding space.

⁴¹ For the given sample, we assume that Americans see *hockey* in the sense of *ice-hockey*.

5 Additional SNLI test-set

In the previous section we have identified several lexical inferences that have not been captured by the model and thus mispredicted. However we could only consider a small subset of misclassified samples, since a small lexical overlap would result in multiple interpretations of what the model understands and what not. In fact, even when identifying missing knowledge in our analysis we needed to rely on intuition rather than certainty when assigning mislabelled samples into categories. We also take a look about what the model seemingly knows, based on its correct predictions, shown in Table 5. In the first example humans see, that both sentences describe the same scenario with only the main

Premise/Hypothesis	Label
A young boy wearing a jacket pushing a hand mower on the grass. A girl is mowing the grass.	<i>contradiction</i>
A man is doing a cannon ball into a pool, stadium chairs fill the background. Someone is jumping into water.	<i>entailment</i>
A woman testing a comfortable pillow. The woman's head is in contact with the pillow.	<i>entailment</i>

Table 10: Correctly classified examples.

actor changing. Aiming for NLI and thus NLU, the model should proceed similarly. It cannot be said, whether the model identifies the paraphrasing of *mowing*. Yet, we can see that the only required information for the correct label is the gender of the mowing person. This is, as we have seen in Section §4, a very important feature within the model. In the second sentence, a human knows, that *doing a cannon ball* is a special form of *jumping*. However, a simple heuristic that *h* is entailed by *p* if describes a more general scenario, would be sufficient if the model is able to identify that a *man* is a *person* and *pool* is similar to *water*. Even the alignment to *jumping* and *ball* may be given in the representation, since the model mixed words in sport/activity dimensions. Similarly the last sentence pair, we intuitively find it hard to believe, the process of *putting the head in contact with the pillow* is implied when *testing* it. Again it seems more likely that the high overlap (and potentially the semantic relatedness of “head” and “pillow”) are causing the correct prediction rather than actual NLU.

5.1 Goal of the new test-set

We believe that the high accuracy on SNLI stems from exploiting these simple heuristics, rather than actually encoding the correct meaning of the sentences. We create a new additional test-set (Glockner et al., 2018) for SNLI with adversarial sentence pairs that only differ in one aspect, defined by lexical semantic relations. This will help in three aspects:

- We show that even state-of-the-art models fail to capture simple lexical inferences and a high performance on SNLI is not sufficient evidence for a proper NLU, being heavily dependant on dataset specific patterns. This is motivated by Jia and Liang (2017), who find similar issues in the field of reading comprehension.
- Differing in only one specific aspect enables very accurate estimation whether the model has enough understanding capabilities for the particular relation or not, as we exclude any noise.
- Only measuring the capability for lexical inferences, that are available in a variety of lexical resources, like WordNet, we show the need to incorporate such knowledge bases into neural networks and provide a dataset, to measure the effectiveness of these approaches.

Our claim, that state-of-the-art results highly overestimate the actual NLU capabilities of the model as they rely on patterns within the dataset, is in line with other works, that tackle the problem from different perspectives. (Gururangan et al., 2018) refer to those patterns as “Annotation Artifacts”, arising from similar strategies used by the annotators, when creating the hypothesis for each label. As opposed to our approach, they do not create new samples to reduce the impact of these patterns. Instead, they identify samples that contain enough information solely in their hypothesis, to be classified correctly. After removing these samples, the remaining dataset shows, that state-of-the-art models perform significantly worse. (Dasgupta et al., 2018) focus on the compositional aspect by automatically generating sentences from SNLI by rearranging noun phrases in any order around the words “<NP>[not] more/less <NP>”, thus requiring the model to consider the sentence structure. Based on their results they claim that the high performance on SNLI arises from the fact that word-overlaps or specific single word relations are often sufficient.

While models relying only on external information from distributed word representations achieve strong results, they still depend on the information encoded within these. Especially for mutually exclusive words that appear in similar contexts, this information alone might not only be insufficient but also misleading.

aim for need of ext. res even though snli large + test for nlu based on lexical inference capabilities in line with 2 other papers, same goal, tackle from different directions

herleitung von error analysis: nur wenig analysiert, da schwer wenn größerer overlap auch bei korrekten eher nicht wegen verständnis sondern wegen was anderem

max: In this paper we show that state-of-the-art NLI systems with only pretrained word embeddings as external information trained on any of these datasets are limited in their generalization ability and fail to capture simple inferences. To show that the SNLI test set alone is not a sufficient measure of the language understanding capabilities of a model, we create a new testset, based solely on sentences from the train set as premise and slightly adapted versions of the same sentence as hypothesis. The hypothesis are created by replacing words from several groups of world knowledge that also occur within the SNLI train set. To target the major drawback of distributed word representations we mostly create contradictory samples by replacing mutually exclusive words that generally appear in similar contexts. As the new dataset consists of samples in which the hypothesis only differs in one aspect, this gives deeper insights in what the model does not know and enables an evaluation based on single replacements rather than one single accuracy figure. We show that even though SNLI is considered a very large dataset for NLI it still does not capture enough information to enable proper reasoning based on text understanding across all evaluated semantic categories without additional knowledge. We also show that less frequent appearing words are very challenging for models and that they struggle to differentiate between mutually exclusive words with similar word representations when not enough evidence is within the data. As all premises given in the new dataset are known for each model in identical form and all adapted word changes are also known from the train data, we expect a model that predicts based on proper language understanding to be able to achieve a reasonable performance on the new data set. Based on the significant performance drop in overall results of the models on the new testset, we show that the high performance scores on SNLI testset do not arise from a proper language understanding.

max: We also show that even though SNLI is considered a very large data set for NLI, it still does not capture enough information to enable proper reasoning based on text understanding.

max: As all premises given in the new dataset are known for each model in identical form and all adapted word changes are also known from the train data, we expect a model that predicts based on proper language understanding to be able to achieve a reasonable performance on the new data set.

5.2 Dataset

We now describe how we create the new test-set and make sure it, that is correct and *fair* w.r.t. the train data in the way that it does not introduce new information, but only relies on the generalization ability. This is important, since we cannot assume a model trained on a specific dataset to perform equally on different domains (Goldberg, 2017).

5.2.1 Creation of adversarial samples

We derive all new sentence pairs from the original SNLI train set, by replacing selected words⁴² within a single sentence. The original sentence is kept as the premise while the adapted sentence serves as the hypothesis. We will in the remainder of this section refer to w_p for the word within p that was replaced by the word w_h in h . We show samples from the resulting samples in Table 5.2.1. Note that traditional RTE systems would consider the first example as neutral, since a

Premise/Hypothesis	Label
The man is holding a saxophone The man is holding an electric guitar	contradiction
A little girl is very sad. A little girl is very unhappy.	entailment
A couple drinking wine A couple drinking champagne	neutral

Table 11: Examples from the newly generated test set.

man may hold both instruments at the same time. However for being conform with the labelling scheme in SNLI this is

⁴² We also allow some multi-word expressions like *New Zealand*.

considered as contradicting⁴³, based on the event-coreference assumption and the most dominant aspects of the image described within the sentence, which was introduced by Bowman et al. (2015) for exactly this purpose of distinguishing between different interpretations.

Generation of word-pairs

We manually generate a list of word-pairs (w_p, w_h) from online resources for English Learning⁴⁴. They provide large lists topically clustered words, which we use to derive co-hyponym, as well collections of as rather generally applicable synonyms or antonyms. We focus partly on entailing, but mostly on contradicting examples, and assume synonyms to refer to the former, co-hyponyms and antonyms to the latter case. Doing so we must consider the following things:

- **Compatible co-hyponyms:** Co-hyponyms not necessarily exclude each other (Kruszewski and Baroni, 2015). A jongleur for instance might also be a clown, even though both could be considered as neighbouring hyponyms of artist. Aiming to see whether the model is able to identify contradicting examples and not too heavily biased by the large lexical overlap, due to our generation process, we identify generally mutually exclusive cohyponyms. Similarly, we remove word-pairs, that commonly are confused by humans like “pink” vs. “purple”.
- **Polysemy:** In order to automatically generate new sentences from (w_p, w_h) , we require both words to have one highly dominant sense to be generally replaceable. We verify this, by sampling random sentences of the word being replaced to see their usage within SNLI. Words that appear in highly different senses are excluded. The country *Jordan* for instance is mostly used in the sense of the basketball player *Michael Jordan* and thus is not used. We observe a similar problem on much more fine-grained level. The word-pair of the antonyms (*old*, *young*) both contradict each other on a very general basis, yet whether they can be replaced or not is dependant on the context. While “old” may refer to things as well as people (like “an old computer” or “an old man”), “young” usually can only be used in combination with people. We thus distinguish between $(w_p < - > w_h)$, that can be swapped both ways, and $(w_p < - w_h)$, that only can be swapped in one direction. In this case the more restricted term “young” can be replaced by “old”, not vice versa, as we aim for a high precision on correct sentences.
- **Structural word usage:** Furthermore, (w_p, w_h) may be used together with different function words. Consider the for instance (*day*, *night*) or (*near*, *far*). While both (w_p, w_h) represent opposite meanings with a high frequency, replacing one with each other leads to invalid sentences like “John sleeps at day[/night]” or “The house is very near[/far] from the sea”. We identify these patterns by looking at the word usage, and extend the (w_p, w_h) adequately with the function words like (*during the day*, *at night*), automatically reducing the chance of incompatible senses as a side-effect.

We manually evaluate all selected (w_p, w_h) for synonyms and antonyms based on the points mentioned above. Topic related co-hyponyms are only individually evaluated, as mapping each cohyponym with each of its co-hyponyms is rather inefficient to manually verify. In addition to that we create antonym word-pairs by from WordNet, sharing the same POS and having a cosine similarity of $\geq XX$. In total we generate 3990 word-pairs⁴⁵ and keep the topical (in case of cohyponyms) or relational (for synonyms and antonyms) information, leading to 13 groups⁴⁶. To ensure that we not confront the models, trained on SNLI, we ensure that each word indeed is within the train-data⁴⁷ and the used word-embeddings. In the final test-set, frequencies⁴⁸ from newly introduced words w_h range from a single occurrence (e.g. “Portugal”) up to 248,051 occurrences (“man”) with a mean of 3,663.1 and a median of 149.5 (IQR = XXX). As the general goal of machine learning is not to memorize we aim for this distribution, havin a high amount of less representative words to measure the models generalization power, if only distributed information from the word-embeddings is given as additional knowledge towards the plain data. We motivate this, as it is very likely in a real-world scenario to encounter samples requiring this kind of knowledge, even though it may not be omnipresent within the train data.

Generation of sentence-pairs

Not wanting to introduce new information that is not present within the original SNLI train-data, we derive all our samples from premises of from the training set. The premises serve in the exact same form as the premise of our new samples, while the hypothesis is generated by replacing w_p within p by w_h . Thus, not only are the newly introduced words known from the training process, but also each p has been seen during training in the exact same form and has

⁴³ We verified this by finding similar samples within the actual SNLI dataset, also labelled as contradiction.

⁴⁴ TODO

⁴⁵ we count the replacement directions, thus $w_p < - > w_h$ counts as $w_p - > w_h$ and $w_p < - w_h$.

⁴⁶ Countries, Nationalities, Colors, Numbers, Antonyms, Synonyms, Vegetables, Drinks, Movements, Materials, Planets, Rooms, Instruments

⁴⁷ Using RegExpr: “YYb<word>YYb”

⁴⁸ The amount of individual sentences containing w_h in the exact surface form.

been encoded with respect to at least three hypothesis, for each label respectively. By doing so we intentionally violate common practices in machine learning as the test-set is not completely isolated from the train data, which should serve as an advantage for the model. We finally remove highly unlikely sentences by ensuring that the bigrams (w^{t-1}, w_h^t) and (w_h^t, w^{t+1}) with w^t being the t th word within h must have a frequency of at least 10 in the wikipedia bigram corpus⁴⁹. In case the replacement consists of several words, w_h^t corresponds to the first or last word of this expression respectively. This preprocessing helps to clean the created samples, yet two problems remain:

Remaining issues

Especially on the semantic level, our newly created sentences may still be incorrect. For instance consider the following sentence-pair:

Premise: The car would not *start* and, consequently, stayed in this garage.

Hypothesis: The car would not *end* and, consequently, stayed in this garage.

Obviously both bigrams (“not”, “end”) and (“end”, “and”) are relatively usual. However does “end” neither serve as an appropriate verb for “car”, nor would the resulting sentence (even if a more applicable word like “stop” would be used) make any sense to a human.

Furthermore, we need to assign the correct label, indicating the relationship between both sentences. Knowing that our newly created p and h only differ in one word, one could consequently assume that the relationship between p and h is the same as the relation between w_p and w_h , thus for instance labelling (p, h) as contradiction, if w_p and w_h exclude each other or as entailment, if they have synonym meanings. This heuristic may indeed be correct for most cases, MacCartney and Manning (2007) however show that this is only the case for upward monotone sentences. In upward monotone sentences, replacing a word (e.g. “cow”) with a more general term, like a hypernym (e.g. “animal”), yields in a broader meaning coverage of the sentence and thus results in entailment. MacCartney and Manning (2007) identify several linguistic patterns like negation or restrictive quantifiers (e.g. “without”) or verbs (e.g. “fail”) that result in downward-monotone sentences, yielding to different results w.r.t. entailment relation. This differentiation is not only relevant for hypernyms but also for other lexical relations, as shown in Table 5.2.1 with the example of co-hyponyms. In both examples “France”

	Sentences	Label
Upward monotone	John is hiking in <i>France</i>	contradiction
	John is hiking in <i>Italy</i>	
Downward monotone	John is hiking outside of <i>France</i>	neutral
	John is hiking outside of <i>Italy</i>	

Table 12: Comparison of co-hyponyms in upward-monotone and downward-monotone sentences.

is replaced by its co-hyponym “Italy”. Even though both words are mutually exclusive only for the upward-monotone sentences, the sentence relation reflects the relation between both words. The same implication does not hold anymore for the downward-monotone sentence. Even though we can assume that most sentences are upward monotone, especially since image captions are more likely to explicitly describe the content of the picture, we must take into account, that whether the sentence relation corresponds to the relation of (w_p, w_h) or not, depends on the context of these words.

5.2.2 Validation

We address both mentioned problems by annotating the new testset using crowd-sourcing with Amazon Mechanical Turk⁵⁰. In order to make this a cost-effective process, we aim for the Human Intelligence Task (HIT) to be simple for the annotators while at the same time validating and labelling as many as possible samples. We constrain ourselves such that one HIT contains five hypothesis, which are all originating from the same premise. This way annotators must only read the premise once, to compare it with those newly created sentences. Not all our identified categories of word-pairs are well represented. In order to get the most of our less frequent categories, we sample 10,000 sentence-pairs to be annotated in a greedy manner: After sorting categories by the amount of samples we created for each of them, we start with the least representative categories and sample as many sentence-pairs as possible, keeping our constraint of needing a multiple of five hypothesis per premise up to an upper bound of samples per category. If more samples are required to complete a HIT with five hypothesis, the next categories are checked in the order of representativeness. Furthermore we keep track of the amount of word-pairs that are included in the sampled sentence-pairs and always prefer less-frequent word-pairs, if several options exist.

⁴⁹ <https://github.com/rmaestre/Wikipedia-Bigram-Open-Datasets>

⁵⁰ <https://www.mturk.com/>

Annotation process

To simplify the HITs for annotators, such that they do not need to understand the labelling scheme of SNLI, we create a set of questions, highly aligned⁵¹ with those proposed of Bowman et al. (2015) that we later map to entailment labels. Specifically, we ask:

1. if both sentences describe **the same event**
2. if the hypothesis **adds new information**
3. if the sentence is **invalid**

Samples that are answered negatively for (1) result in the label *contradiction*. If (1) is answered positively, the label is either *neutral*, if (2) is answered positively, or *entailment*, if (2) is answered negatively. Sentences that show major grammatical errors or do not make sense to a native english speaker should be marked using (3) and no label is inferred. We defined the HIT user interface such that no other combinations can be selected as an answer. The questions are explained in deeper detail in an additional introduction and we ensure they are understood correctly with a mandatory qualification test. Since SNLI does contain grammatical and spelling errors, we specifically allowed minor errors of that kind. While SNLI does contain fictive sentences that are not realistic, it is hard to define the degree that sentences are allowed to sound unrealistic. While someone “flying to school” instead of “walking to school” may be a bit unrealistic but still plausible for a fictive scenario, “sitting at the table and eating” and “walking at the table and eating” seems indeed very unlikely. Yet, this is highly dependant on the subjective perspective. We defined this aspect of (3) rather swammy, by allowing fictive scenarios, however counting on the English capabilities of native speakers to identify if those sentences make sense to them and seem like a proper usage of words. Due to this loose definition, we strictly ignore all samples, if at least one single annotator marked them as invalid. We assure that our workers have shown to annotate appropriate to the task description, we only accept annotators with an approval rate of at least 99% and a minimum of 1,000 prior tasks. A sample HIT is shown in Figure 23. The questions are explained to the users with a sample in the same form in

Original sentence		Describes same event	Adds information	Not correct/ grammatical
People are at a stadium on a nice day waiting for the football game to start.				
New sentence				
People are at a stadium on a nice day waiting for the football game to stop .	<input type="radio"/> Yes <input type="radio"/> No	<input type="checkbox"/>	<input type="checkbox"/>	
People are at a stadium on a nice night waiting for the football game to start.	<input type="radio"/> Yes <input type="radio"/> No	<input type="checkbox"/>	<input type="checkbox"/>	
People are at a stadium on a nice day waiting for the football game to end .	<input type="radio"/> Yes <input type="radio"/> No	<input type="checkbox"/>	<input type="checkbox"/>	
People are at a stadium on a nice day waiting for the football game to finish .	<input type="radio"/> Yes <input type="radio"/> No	<input type="checkbox"/>	<input type="checkbox"/>	
People are at a stadium on an awful day waiting for the football game to start.	<input type="radio"/> Yes <input type="radio"/> No	<input type="checkbox"/>	<input type="checkbox"/>	

Figure 23: Example of a HIT in Amazon Mechanical Turk.

the instructions. In order to make the task more attractive to annotators without increasing the payment too much, we simplified the process by highlighting the changed words.

We assign each HIT to three annotators. After removing all invalid sentence-pairs, we consider the majority label as the gold label, if at least two annotators agree on it. Sentence-pairs without agreement are not considered for our new testset. After an initial annotation round of 1,000 samples, we remove categories, that show a high tendency of having invalid sentences. We show the statistics of our final testset, consisting of 8,193 sentence-pairs in Table 13. As can be seen, the new testset is heavily focused on contradicting samples. Since this testset only serves to measure the capabilities of a model to deal with lexical inferences and not to replace the original SNLI testset, this is not problematic. It still must

⁵¹ Bowman et al. (2015) only provide their annotation guidelines for the task of creating new hypothesis, not for validating them. The validation task was conducted separately and was only open for annotators who participated in the hypothesis-generation task and thus were qualified already.

	Instances				Fleiss κ		
	contradiction	neutral	entailment	Overall	contradiction	entailment	Overall
SNLI Test	3,236	3,215	3,364	9,815	0.77	0.69	0.67
New Test	7,164	47	982	8,193	0.61	0.90	0.61

Table 13: Statistics of SNLI testset compared with the newly generated testset.

be considered when evaluating a model, as a simple baseline predicting everything as contradiction would result in an accuracy of 87.4%. We report the agreement with Fleiss Kappa (?) (over all samples and for the representative labels *entailment* and *contradiction*), as it was done by Bowman et al. (2015) for the original SNLI. For better comparison we re-calculate the numbers on all valid samples of the SNLI testset. The new testset yields “substantial agreement” with a Fleiss Kappa of 0.61. It should be noted that Kappa also considers how likely a specific label would be selected by chance, and for this purpose takes the overall label distribution into account. This does not influence the measure for the original SNLI, since all labels are evenly distributed. For the new testset however, this method assumes that contradiction is more likely to appear by chance, due to its high frequency. As this results from our selected word-pairs rather than an underlying “natural distribution”, this Fleiss Kappa might be less suited as an agreement measure for the new dataset. This being said, it indeed is likely that annotators are slightly biased towards selecting contradiction, as a result of our HIT presentation. As most word-pairs are contradicting and word differences are highlighted, annotators might to some extent shift their focus more on the difference between those words, than on the actual word usage in context. To prove an easier to interpret measure of this dataset, we estimate the human performance in the same manner as done by Gong et al. (2017) for SNLI. They considering all samples with majority label, resulting in the gold label, and calculate the ratio of annotator labels matching the majority gold label as an estimate for the human performance in terms of accuracy. Thus, let $g(x, y)$ with x as the annotator label and y as the estimated gold label define if the annotation counts as a misclassification or not:

$$g(x, y) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases} \quad (9)$$

Let furthermore L contain all pairs (x, y) of the annotated dataset and $|L|$ be the amount of elements within L . Following Gong et al. (2017), we estimate the human performance a in accuracy as follows:

$$a = \frac{\sum_{(x,y) \in L} g(x, y)}{|L|} \quad (10)$$

Doing so, we estimate the human performance on the new testset to be 94.1%, slightly higher than the human performance estimated on SNLI with only 87.7%, indicating that our new sentence pairs do not pose additional difficulties, but in fact indeed seem relatively easy for humans.

5.3 Evaluation

We evaluate three neural models without external information other than distributed word-embeddings that achieve strong results on SNLI. All models are retrained using the provided code, keeping all hyperparameters, with different datasets. We explain the experimental setup and results below.

5.3.1 Experimental setup

Models without external knowledge

We evaluate the Residual-Stacked Encoder⁵² (Nie and Bansal, 2017), as explained in Section §2.3, ESIM (Chen et al., 2017b) and Decomposable Attention (Parikh et al., 2016), both explained in Section §3.3. The published model of ESIM ensembles two models with different sentence encoding strategies, one is based on a TreeLSTM, the other on a biLSTM. For our experiments we retrain only the biLSTM-based model. For Decomposable-Attention we use the AllenNLP re-implementation⁵³. As opposed to the reported version on the SNLI leaderboard⁵³ this implementation does not use the optional intra-sentence attention. Its performance on the SNLI test is with 84.7% slightly lower but comparable to the model with intra-sentence attention (86.3%). All models have different characteristics, depicted in Table 5.3.1 and are at the time of the experiment amongst the best within their categories.

⁵² <http://allennlp.org/models>

⁵³ <https://nlp.stanford.edu/projects/snli/>

	Finetune Embeddings	LSTM-based	Inter-sentence Attention
Decomposable Attention (Parikh et al., 2016)	—	—	yes
Residual-Stacked Encoder Nie and Bansal (2017)	yes	yes	—
ESIM (Chen et al., 2017b)	yes	yes	yes

Table 14: Architectural comparison of tested neural models without external knowledge.

5.3.2 Models with external knowledge

Additionally we also provide a simple WordNet baseline, predicting the relationship of (p, h) by assuming all sentences to be upward monotone and thus have the same relation as (w_p, w_h) . Specifically, for each (w_p, w_h) we check their lexical relation within WordNet and map it to a relation label in the following manner:

- **Synonymy:** Synonyms are predicted as *entailment*.
- **Anotnomy:** Antonyms are predicted as *contradiction*.
- **Hypernymy:** If w_p is a hypernym of w_h the sentence-pair is predicted as *neutral*, if w_p is a hyponym of w_h as *entailment*.
- **Co-hyponymy:** Cohyponyms are predicted as *contradiction*. We only consider co-hyponyms with a maximum distance to their common hypernym of two edges, as considering all potential co-hyponyms would yield all (w_p, w_h) coming from the same root word to be labelled as contradiction.

We mapped multi-word expressions like “at night” to their meaning carrying word (“night”), if the function words have only been added for a higher precision when replacing the words in the generation process. In case they refer to actual entities (e.g. “New Zealand”) we identify the applicable synsets of the whole expression in WordNet. As explained in Section §3.1.1, words may have several synsets leading to potentially several different lexical relations between the words of interest w_p and w_h . We calculate for each relation between both words a score $s = \max(r_p, r_h)$ with r_p and r_h being the rank of the synset of the word w_p and w_h respectively. This follows the common heuristic that dominant senses appear as the first synsets while rare senses appear at the end. Subsequently, if several lexical relations exist, we consider the one with the lowest assigned score as tie-breaker⁵⁴ and thus tend to focus on more dominant word-senses. Of course, this baseline only is possible if knowing that p and h only differ in w_p and w_h and is thus not applicable to sentence-pairs in general. Yet it provides insight to what extend information within WordNet can help on our new testset. In addition to that we also report⁵⁵ the results of KIM (Chen et al., 2017a), as explained in Section §3.3.2.

Training data

In addition to training all models on SNLI, which is considered relatively easy, we also train each model on the union of SNLI and MultiNLI and SciTail respectively, both are assumed to be more difficult and explained in deeper detail in Section §3.2. The motivation is that while the SNLI might lack the training data needed to learn the required lexical knowledge, it may be available in the other datasets, which are presumably less simple.

5.3.3 Results

The results for each model with the according train sets are visualized in Table 5.3.3. There is a clear trend that adding MultiNLI to the training data boosts the model’s performance on the new test set. At the same time it decreases the test accuracy on SNLI, indicating that the performance gained on SNLI test does not reflect the true NLU capabilities, since clearly less lexical semantic relations are understood. Yet, compared with the original estimated performance even by almost doubling the amount of train data with MultiNLI all models without external knowledge show a significant drop in performance. While MultiNLI follows the same labelling scheme as SNLI and thus is compatible, SciTail does not specifically assume event coference and lacks having the label *contradiction* which is dominant in the new test set. Thus the models seem to not be able to leverage from the extended amount of data in this case. Both models with WordNet information perform significantly higher than the ones without. This shows, that the only lexical relations as contained in WordNet are sufficient to gain strong improvements on the new dataset. In addition to that, especially since apparently

⁵⁴ In case the score s is identical for several relations, we select the relation in the following order:
synonym > antonym > hypernym > hyponym > co-hyponym

⁵⁵ We did not conduct the experiment with KIM ourselves, but received their results from the original authors recently. The analysis part thus does not contain deeper analysis of the performance of KIM.

Model	Train set	SNLI test set	New test set	Δ
Decomposable Attention (Parikh et al., 2016)	SNLI	84.7%	51.9%	-32.8
	MultiNLI + SNLI	84.9%	65.8%	-19.1
	SciTail + SNLI	85.0%	49.0%	-36.0
ESIM (Chen et al., 2017b)	SNLI	87.9%	65.6%	-22.3
	MultiNLI + SNLI	86.3%	74.9%	-11.4
	SciTail + SNLI	88.3%	67.7%	-20.6
Residual-Stacked Encoder [◇] (Nie and Bansal, 2017)	SNLI	86.0%	62.2%	-23.8
	MultiNLI + SNLI	84.6%	68.2%	-16.8
	SciTail + SNLI	85.0%	60.1%	-24.9
WordNet Baseline	-	-	85.8%	-
KIM (Chen et al., 2017a)	SNLI	88.6%	83.5%	-5.1

Table 15: Results of models on the new test set compared with the original SNLI test set.

those relations have also shown to be useful for KIM in training and thus are considered for the prediction, it shows some crucial drawbacks in the neural models without WordNet, as they clearly lack to extract those features when trained solely with textual input. Only dropping by 5.1 points in accuracy with respect to the performance on SNLI, KIM seems substantially more stable, when sentences are adapted, forcing the model to predict based on a language understanding rather than dataset specific patterns. This of course may be due to the fact that the new testset, being closely related to the train data and requiring the same additional knowledge as added for KIM, is highly suitable for KIM. Thus, it still does not prove truly superior NLU, yet it shows that they found a successful strategy of integrating this resource into neural models.

5.4 Analysis

We take a closer look on the performance achieved by the models without external knowledge. As the performance only improved marginally, if the train data is tremendously increased by adding MultiNLI we focus our analysis part on the models solely trained on the original SNLI dataset.

5.4.1 Accuracy by category

Table 5.4.1 shows the accuracy per category, as defined when creating the word-pairs, for all models, including KIM and the WordNet baseline. As not all categories contain an even amount of samples, we additionally supply this information, together with sample words for a better understanding what each category represents. Originating from our word-pairs, almost all categories are majority labelled contradiction, solely synonyms mostly are labelled as entailment. All neural models achieve good results on categories that occur very frequently within SNLI in general, like *colors*. Also *instruments* are well captured. We find that musical instruments often occur in SNLI in similar sentences containing some kind of actor and in conjunction with the instrument and the verbs "hold" and especially "play". As in contradicting samples in most cases the instrument changes, our newly created sentence-pairs are very similar to those within the train data. On the other hand, categories that are rare in SNLI, like *Planets*, are not well understood by those models. As opposed to *instruments*, the relevance of *ordinals* within a sentence in SNLI usually is less crucial. Yet this originates only from the commonly applied strategies by the annotators as identified by Gururangan et al. (2018). Consider for instance the sentence "A man racing his motorcycle comes in *first*.", which naturally yields in contradiction if we replace "first" by "eights". Yet the model without external information seemingly do not differentiate between both ordinal numbers, as annotators rather tend to change "man" or even "motorcycle". Also *drinks*, *vegetables* and *rooms* are generally harder for the model to predict. The reason for *antonyms* being more difficult than *antonyms_WordNet* most likely arises from the fact that they include gender-related antonyms, appearing in SNLI in abundance, that we do not include within the handcrafted word pairs. As *synonyms* examples have large lexical overlap and differ only by one word, occurring in similar contexts and subsequently having a similar word-vector, it is no surprise, that all models achieve a good performance here. Inter-sentence-attention seems to have an advantage in this case, since only the Residual Encoders does not improve significantly over its original test performance. We thus focus the remaining part of the analysis on contradicting sentence-pairs only.

Category	Amount	Example Words	Decomposable Attention	ESIM	Residual Encoders	WordNet Baseline	KIM
antonyms	1,147	<i>loves - dislikes</i>	41.6%	70.4%	58.2%	95.5%	86.5%
cardinals	759	<i>five - seven</i>	53.5%	75.5%	53.1%	98.6%	93.4%
nationalities	755	<i>Greek - Italian</i>	37.5%	35.9%	70.9%	78.5%	73.5%
drinks	731	<i>lemonade - beer</i>	52.9%	63.7%	52.0%	94.8%	96.6%
antonyms (WN)	706	<i>sitting - standing</i>	55.1%	74.6%	67.9%	94.5%	78.8%
colors	699	<i>red - blue</i>	85.0%	96.1%	87.0%	98.7%	98.3%
ordinals	663	<i>fifth - 16th</i>	2.1%	21.0%	5.4%	40.7%	56.6%
countries	613	<i>Mexico - Peru</i>	15.2%	25.4%	66.2%	100.0%	70.8%
rooms	595	<i>kitchen - bathroom</i>	59.2%	69.4%	63.4%	89.9%	77.6%
materials	397	<i>stone - glass</i>	65.2%	89.7%	79.9%	75.3%	98.7%
vegetables	109	<i>tomato - potato</i>	43.1%	31.2%	37.6%	86.2%	79.8%
instruments	65	<i>harmonica - harp</i>	96.9%	90.8%	96.9%	67.7%	96.9%
planets	60	<i>Mars - Venus</i>	31.7%	3.3%	21.7%	100.0%	5.0%
synonyms	894	<i>happy - joyful</i>	97.5%	99.7%	86.1%	70.5%	92.1%
total	8,193		51.9%	65.6%	62.2%	85.8%	83.5%

Table 16: Accuracy reached for the tested models for each category with assoziated sample words and the amount of instances.

5.4.2 Impact on the word embeddings

As words sharing both lexical relations, antonymy and co-hyponymy, in many cases result in similar word representations with distributed embeddings, we analyse the impact of those representations. For this, we leverage the fact that our new dataset only differs in one word.

Without finetuning

Figure 24 visualizes the performance of all contradicting samples, compared to the cosine similarity between the word-vectors of w_p and w_h achieved by Decomposable Attention. We exclude any multi-word expressions in this analysis. Let v_p

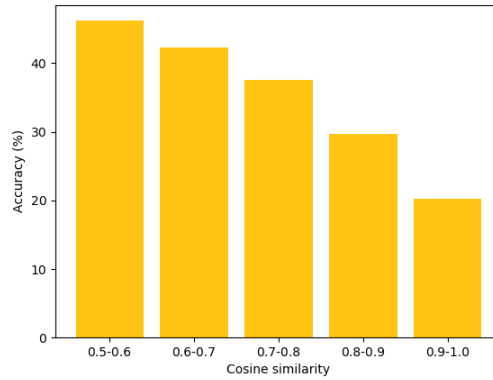


Figure 24: TODO.

and v_h be the word vectors of the GloVe embeddings used by Decomposable Attention, the according cosine similarity $\cos(v_p, v_h)$ is calculated as follows:

$$\cos(v_p, v_h) = \frac{v_p \cdot v_h}{\|v_p\| \|v_h\|} \quad (11)$$

- embeddings - embeddings finetune max: The performance gain on MultiNLI+SNLI most likely arises from the fact that almost 50% of the replacement words used to generate the new dataset occur in MultiNLI in a way 612/1277 phrase replacement pairs (p_p, p_h) occur in the MultiNLI training data in contradiction pairs such that p_p is in the premise and p_h is in the hypothesis. -words esim -wp

6 Approaches to incorporate WordNet information

6.1 Extraction of WordNet data

6.2 Integrating information into word-embeddings

6.2.1 Motivation

(Rubinstein et al., 2015) show something that distributional embeddings not always good (reread)

6.2.2 Concatenating pre-trained word-embeddings

6.2.3 Concatenation categorical information

6.2.4 Analysis

6.3 Multitask Learning

(Levy et al., 2015) show that embeddings not necessarily (reread)

6.3.1 Motivation

6.3.2 Architecture

6.3.3 Approaches

6.3.3.1 Different sizes of multitask MLP

6.3.3.2 Introducing Dropout

6.3.3.3 Introducing an additional shared layer

6.3.3.4 Fixing multitasking network during training

6.3.3.5 Focusing on original words within sentence representation

6.3.4 Analysis

6.3.5 Evaluation

cite: An Overview of Multi-Task Learning in Deep Neural Networks

7 Conclusion

References

- Jorge A Balazs, Edison Marrese-Taylor, Pablo Loyola, and Yutaka Matsuo. 2017. Refining raw sentence representations for textual entailment recognition via attention. *arXiv preprint arXiv:1707.03103*.
- Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.
- Christopher M. Bishop. 2007. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer.
- Johan Bos and Katja Markert. 2005. Recognising textual entailment with logical inference. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 628–635. Association for Computational Linguistics.
- Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.
- Samuel R Bowman, Jon Gauthier, Abhinav Rastogi, Raghav Gupta, Christopher D Manning, and Christopher Potts. 2016. A fast unified model for parsing and sentence understanding. *arXiv preprint arXiv:1603.06021*.
- Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1994. Signature verification using a "siamese" time delay neural network. In *Advances in Neural Information Processing Systems*, pages 737–744.
- Asli Celikyilmaz, Dilek Hakkani-Tur, Panupong Pasupat, and Ruhi Sarikaya. 2010. Enriching word embeddings using knowledge graph for semantic tagging in conversational dialog systems. *genre*.
- Stergios Chatzikyriakidis, Robin Cooper, Simon Dobnik, and Staffan Larsson. 2017. An overview of natural language inference data collection: The way forward? In *Proceedings of the Computing Natural Language Inference Workshop*.
- Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, and Diana Inkpen. 2017a. Natural language inference with external knowledge. *arXiv preprint arXiv:1711.04289*.
- Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017b. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1657–1668.
- Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017c. Recurrent neural network-based sentence encoder with gated attention for natural language inference. *arXiv preprint arXiv:1708.01353*.
- Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733*.
- Peter Clark, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter D Turney, and Daniel Khashabi. 2016. Combining retrieval, statistics, and inference to answer elementary science questions. In *AAAI*, pages 2580–2586.
- Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Johan Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, et al. 1996. Using the framework. Technical report, Technical Report LRE 62-051 D-16, The FraCaS Consortium.
- Ido Dagan. 2000. Contextual word similarity. *Handbook of Natural Language Processing*, pages 459–475.
- Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, 15(4):i–xvii.
- 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 textual entailment*, pages 177–190. Springer.
- Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J Gershman, and Noah D Goodman. 2018. Evaluating compositionality in sentence embeddings. *arXiv preprint arXiv:1802.04302*.
- Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615.

-
- Reza Ghaeini, Sadid A Hasan, Vivek Datla, Joey Liu, Kathy Lee, Ashequl Qadir, Yuan Ling, Aaditya Prakash, Xiaoli Z Fern, and Oladimeji Farri. 2018. Dr-bilstm: Dependent reading bidirectional lstm for natural language inference. *arXiv preprint arXiv:1802.05577*.
- Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking nli systems with sentences that require simple lexical inferences. In *The 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.
- Yoav Goldberg. 2017. *Neural Network Methods in Natural Language Processing (Synthesis Lectures on Human Language Technologies, Band 37)*. Morgan & Claypool Publishers.
- Yichen Gong, Heng Luo, and Jian Zhang. 2017. Natural language inference over interaction space. *arXiv preprint arXiv:1709.04348*.
- Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610.
- Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M Meyer, and Christian Wirth. 2012. Uby: A large-scale unified lexical-semantic resource based on lmf. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 580–590. Association for Computational Linguistics.
- Iryna Gurevych, Judith Eckle-Kohler, and Michael Matuschek. 2016. Linked lexical knowledge bases: Foundations and applications. *Synthesis Lectures on Human Language Technologies*, 9(3):1–146.
- 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*.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Zhiting Hu, Zichao Yang, Ruslan Salakhutdinov, and Eric Xing. 2016. Deep neural networks with massive learned knowledge. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1670–1679.
- Nancy Ide and Catherine Macleod. 2001. The american national corpus: A standardized resource of american english. In *Proceedings of Corpus Linguistics*, volume 3, pages 1–7. Lancaster University Centre for Computer Corpus Research on Language Lancaster, UK.
- Nancy Ide and Keith Suderman. 2004. The american national corpus first release. In *LREC*.
- Nancy Ide and Keith Suderman. 2006. Integrating linguistic resources: The american national corpus model. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*. Citeseer.
- Jinbae Im and Sungzoon Cho. 2017. Distance-based self-attention network for natural language inference. *arXiv preprint arXiv:1712.02047*.
- Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.
- Daniel Jurafsky and James H. Martin. 2008. *Speech and Language Processing, 2nd Edition*. Prentice Hall.
- Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTail: A textual entailment dataset from science question answering. In *AAAI*.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Germán Kruszewski and Marco Baroni. 2015. So similar and yet incompatible: Toward the automated identification of semantically compatible words. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 964–969.
- Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436.
- Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

-
- Quan Liu, Hui Jiang, Si Wei, Zhen-Hua Ling, and Yu Hu. 2015. Learning semantic word embeddings based on ordinal knowledge constraints. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1501–1511.
- Bill MacCartney, Michel Galley, and Christopher D Manning. 2008. A phrase-based alignment model for natural language inference. In *Proceedings of the conference on empirical methods in natural language processing*, pages 802–811. Association for Computational Linguistics.
- Bill MacCartney and Christopher D Manning. 2007. Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 193–200. Association for Computational Linguistics.
- Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.
- Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 1–8.
- Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Using automatically acquired predominant senses for word sense disambiguation. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
- Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751.
- George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.
- Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2015. Natural language inference by tree-based convolution and heuristic matching. *arXiv preprint arXiv:1512.08422*.
- Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association of Computational Linguistics*, 5(1):309–324.
- Tsendsuren Munkhdalai and Hong Yu. 2017. Neural semantic encoders. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 1, page 397. NIH Public Access.
- M Lynne Murphy. 2003. *Semantic relations and the lexicon: Antonymy, synonymy and other paradigms*. Cambridge University Press.
- Nikita Nangia, Adina Williams, Angeliki Lazaridou, and Samuel R Bowman. 2017. The repeval 2017 shared task: Multi-genre natural language inference with sentence representations. *arXiv preprint arXiv:1707.08172*.
- Yixin Nie and Mohit Bansal. 2017. Shortcut-stacked sentence encoders for multi-domain inference. *arXiv preprint arXiv:1708.02312*.
- Patrick Pantel. 2005. Inducing ontological co-occurrence vectors. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 125–132. Association for Computational Linguistics.
- Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*.
- 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)*, pages 1532–1543.
- Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

-
- Rion Snow Sushant Prakash, Daniel Jurafsky, and Andrew Y Ng. 2007. Learning to merge word senses. *EMNLP-CoNLL 2007*, page 1005.
- Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th international joint conference on Artificial intelligence-Volume 1*, pages 448–453. Morgan Kaufmann Publishers Inc.
- Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiský, and Phil Blunsom. 2015. Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664*.
- Dana Rubinstein, Effi Levi, Roy Schwartz, and Ari Rappoport. 2015. How well do distributional models capture different types of semantic knowledge? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 726–730.
- Andreas Rücklé, Steffen Eger, Maxime Peyrard, and Iryna Gurevych. 2018. Concatenated p -mean word embeddings as universal cross-lingual sentence representations. *arXiv preprint arXiv:1803.01400*.
- Magnus Sahlgren. 2008. The distributional hypothesis. *Italian Journal of Disability Studies*, 20:33–53.
- Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Sen Wang, and Chengqi Zhang. 2018. Reinforced self-attention network: a hybrid of hard and soft attention for sequence modeling. *arXiv preprint arXiv:1801.10296*.
- Vered Shwartz, Omer Levy, Ido Dagan, and Jacob Goldberger. 2015. Learning to exploit structured resources for lexical inference. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 175–184.
- Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM.
- Marta Tatu and Dan Moldovan. 2005. A semantic approach to recognizing textual entailment. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 371–378. Association for Computational Linguistics.
- Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2017. A compare-propagate architecture with alignment factorization for natural language inference. *arXiv preprint arXiv:1801.00102*.
- Ivan Vulić and Nikola Mrkšić. 2017. Specialising word vectors for lexical entailment. *arXiv preprint arXiv:1710.06371*.
- Ivan Vulić, Nikola Mrkšić, Roi Reichart, Diarmuid Ó Séaghdha, Steve Young, and Anna Korhonen. 2017. Morph-fitting: Fine-tuning word vector spaces with simple language-specific rules. *arXiv preprint arXiv:1706.00377*.
- Johannes Welbl, Nelson F Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. *arXiv preprint arXiv:1707.06209*.
- Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.
- Chang Xu, Yalong Bai, Jiang Bian, Bin Gao, Gang Wang, Xiaoguang Liu, and Tie-Yan Liu. 2014. Rc-net: A general framework for incorporating knowledge into word representations. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1219–1228. ACM.
- Han Yang, Marta R Costa-jussà, and José AR Fonollosa. 2017. Character-level intra attention network for natural language inference. *arXiv preprint arXiv:1707.07469*.
- 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.
- Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Extracting lexical semantic knowledge from wikipedia and wiktionary. In *LREC*, volume 8, pages 1646–1652.

List of Figures

1	A sample ontology of animals to illustrate the lexical relations <i>Hypernymy</i> and <i>Holonymy</i>	7
2	The architecture of the sentence-encoding component within the Shortcut-Stacked-Encoder, taken from Nie and Bansal (2017).	8
3	Example of different synsets of the lemma “table” (only noun senses) within WordNet, taken from http://wordnetweb.princeton.edu	11
4	General architecture of a RNN (left). Example sentence in an unrolled RNN (right).	19
5	Visualized example of extracting interpretable information of the max-pooled sentence representations with a dimensionality of 3.	20
6	The standard deviation within a dimension of sentence representations (x-axis) by the amount of dimensions with the given standard deviation.	21
7	An extraction of a grid-plot, showing dimensions with the position within the sentence of the word, responsible for the dimensional value.	22
8	An extraction of a grid-plot, showing syntactical information using the POS tag with pre-sorted rows to have a single dominant label.	23
9	An extraction of a grid-plot, gender specific female using only sentences with words of pre-defined wordlists.	24
10	Representation visualitation with respect to genders of dimension 199 (left) and dimension 602 (right).	24
11	Detailed representation visualitation of different terms for human males of dimension 199 (left) and dimension 602 (right).	25
12	Detailed representation visualitation of different terms for human females of dimension 845 (left) and dimension 311 (right).	25
13	Dimension 713 encoding verbs (left) and dimension 2020 encoding adjectives.	29
14	Dimension 757, encoding the subjects (left), and dimension 1840 encoding objects (right) of sentences.	30
15	Word alignments of an entailing sentence pair either by counting all shared dimensions (left) or only dimensions with at least a value of 0.2 (right).	31
16	Visualitation of an entailing sample with applied element-wise multiplication either using the mean (left) or maximum (right) product of all shared dimensions for each word pair.	31
17	Visualitation of a contradicting sample by counting meaningful shared dimensions (left) and meaningful distinct dimensions (right) amongst pairs of words.	32
18	Dimension-wise visualitation of distinct information represented by <i>sitting</i> in the premise and <i>standing</i> in the hypothesis.	33
19	Visualitation of a sample sentence pair with explanatory guides for interpretation.	34
20	Visualitation of 150 sentence pairs (p , h), correctly labelled as entailment.	34
21	Visualization of samples predicted as entailment (left) and neutral (right) after swapping p and h	35
22	Visualitazion of 150 sentence pairs (p , h) correctly labelled as <i>neutral</i> (left) and <i>contradiction</i> (right).	36
23	Example of a HIT in Amazon Mechanical Turk.	43
24	TODO.	47

List of Tables

1	Accuracy in percent of different implementations of the model from Nie and Bansal (2017), achieved on the SNLI dataset compared with human performance.	9
2	Example sentence pairs, taken from SNLI, showing typical sentences within the dataset.	13
3	Example sentence pairs from MultiNLI, taken from RepEval 2017 Shared Task, showing samples of different genres.	15
4	Example sentence pairs from SciTail Task, different premises retrieved for two hypothesis.	16
5	Accuracies achieved on SNLI using $ r $ -dimensional sentence representations of gender-specific dimensions.	26
6	Results in terms of accuracy of inverted gender-specific dimensions on SNLI train and dev set.	27
7	Comparison of samples between their predictions based on the original and gender-inverted sentence representations.	27
8	Misclassified samples with gold label <i>contradiction</i> , predicted as <i>entailment</i>	37
9	Misclassified samples with gold label <i>entailment</i> , predicted as <i>contradiction</i>	38
10	Correctly classified examples.	39
11	Examples from the newly generated test set.	40
12	Comparison of co-hyponyms in upward-monotone and downward-monone sentences.	42
13	Statistics of SNLI testset compared with the newly generated testset.	44
14	Architectural comparison of tested neural models without external knowledge.	45
15	Results of models on the new test set compared with the original SNLI test set.	46
16	Accuracy reached for the tested models for each category with assoziated sample words and the amount of instances.	47