Generative models for natural language inference

DGM4NLP

Miguel Rios
University of Amsterdam

May 12, 2019

Outline

Introduction

- Textual entailment is defined as a **directional relation** between pairs of text expressions, the **T** "Text", and the **H** "Hypothesis".

Introduction

- Textual entailment is defined as a **directional relation** between pairs of text expressions, the **T** "Text", and the **H** "Hypothesis".
- We say that T entails H if the meaning of H can be inferred from the meaning of T, as would typically be interpreted by people.
  T $\rightarrow$ H

Introduction

- Textual entailment is defined as a **directional relation** between pairs of text expressions, the **T** "Text", and the **H** "Hypothesis".
- We say that T entails H if the meaning of H can be inferred from the meaning of T, as would typically be interpreted by people.
  $T \rightarrow H$

> T: The purchase of Houston-based LexCorp by BMI for $2Bn prompted widespread sell-offs by traders as they sought to minimize exposure.
> H: BMI acquired an American company.

Recognising Textual Entailment

- Recognition: identification of a thing or person from **previous** encounters or **knowledge**.

Recognising Textual Entailment

- Recognition: identification of a thing or person from **previous** encounters or **knowledge**.
- Physicians are trained in medicine to recognise and treat a disease.

Recognising Textual Entailment

- RTE Challenge (Dagan and Glickman, 2005), provides the first benchmark.

Recognising Textual Entailment

- RTE Challenge (Dagan and Glickman, 2005), provides the first benchmark.
- Participant methods decide for each entailment pair whether T entails H or not.

Recognising Textual Entailment

- RTE Challenge (Dagan and Glickman, 2005), provides the first benchmark.
- Participant methods decide for each entailment pair whether T entails H or not.
- The annotation used for the entailment decision is **TRUE** if T entails H or **FALSE** otherwise.

Recognising Textual Entailment

- RTE Challenge (Dagan and Glickman, 2005), provides the first benchmark.
- Participant methods decide for each entailment pair whether T entails H or not.
- The annotation used for the entailment decision is **TRUE** if T entails H or **FALSE** otherwise.

RTE can be **framed** as a classification problem, where the entailment relations are the classes, and the RTE benchmark provides the essential evidence to build a **supervised binary classifier** (Dagan et al., 2010)

Applications of Textual Entailment

- RTE has been proposed as a **generic task that captures major semantic inference needs** across natural language processing applications.

Applications of Textual Entailment

- RTE has been proposed as a **generic task that captures major semantic inference needs** across natural language processing applications.
- We can frame natural language processing tasks as recognition.
  Input as T and generated output as H.

Question Answering

- Question Answering system generates as output the best candidate answers. While the top candidate may not be the correct answer, the correct answer is in the set of returned candidates.

Question Answering

- Question Answering system generates as output the best candidate answers. While the top candidate may not be the correct answer, the correct answer is in the set of returned candidates.

T/Q: Arabic, for example, is used densely across North Africa and from the Eastern Mediterranean to the Philippines, as the key language of the Arab world.
H/A: Arabic is the primary language of the Philippines.

Summarisation

- Identifying if a new sentence contains information already by a summary-in-progress (redundancy detection) can be framed as the current summary as T and the new sentence as H.

Summarisation

- Identifying if a new sentence contains information already by a summary-in-progress (redundancy detection) can be framed as the current summary as T and the new sentence as H.

> T/S1: Google and NASA announced a working agreement, Wednesday, that could result in the Internet giant building a complex of up to 1 million square feet on NASA-owned property, adjacent to Moffett Field, near Mountain View.
>
> H/S2: Google may build a campus on NASA property.

Outline

Challenge of RTE

T: The purchase of Houston-based LexCorp by BMI for $2Bn prompted widespread sell-offs by traders as they sought to minimize exposure.
H: BMI acquired an American company.

To recognise **TRUE** entailment relation:

Challenge of RTE

T: The purchase of Houston-based LexCorp by BMI for $2Bn prompted widespread sell-offs by traders as they sought to minimize exposure.
H: BMI acquired an American company.

To recognise **TRUE** entailment relation:

- "company" in the Hypothesis can match "LexCorp",

Challenge of RTE

T: The purchase of Houston-based LexCorp by BMI for $2Bn prompted widespread sell-offs by traders as they sought to minimize exposure.
H: BMI acquired an American company.

To recognise **TRUE** entailment relation:

- "company" in the Hypothesis can match "LexCorp",
- "based in Houston" implies "American",

Challenge of RTE

T: The purchase of Houston-based LexCorp by BMI for $2Bn prompted widespread sell-offs by traders as they sought to minimize exposure.
H: BMI acquired an American company.

To recognise **TRUE** entailment relation:

- “company” in the Hypothesis can match “LexCorp”,
- “based in Houston” implies “American”,
- identify the relation “purchase”,

Challenge of RTE

T: The purchase of Houston-based LexCorp by BMI for $2Bn prompted widespread sell-offs by traders as they sought to minimize exposure.
H: BMI acquired an American company.

To recognise **TRUE** entailment relation:

- "company" in the Hypothesis can match "LexCorp",
- "based in Houston" implies "American",
- identify the relation "purchase",
- determine that "A purchased by B" implies "B acquires A".

Levels of Representation

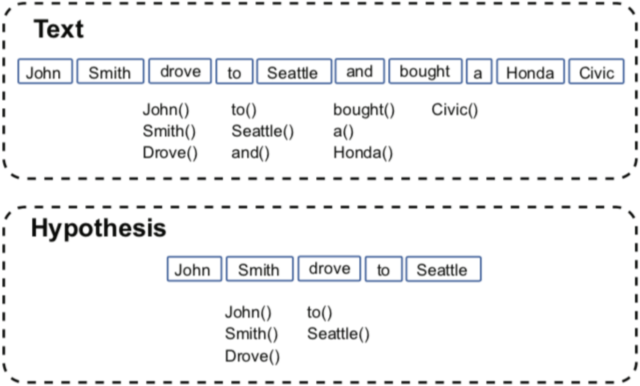
- Determining the equivalence or non-equivalence of the meanings of the T-H.

Levels of Representation

- Determining the equivalence or non-equivalence of the meanings of the T-H.
- The representation (e.g. words, syntax, semantics) of the T-H pair that is used to extract features to train a supervised classifier.

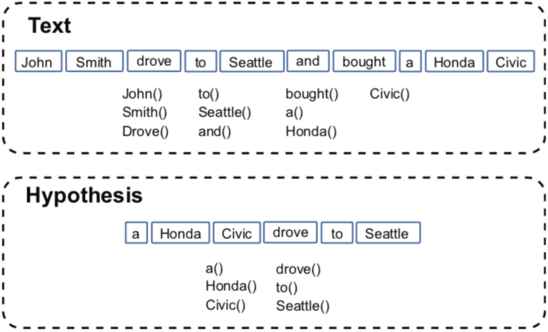
Lexical level

- Every assertion (word) in the representation of H is contained in the representation T.



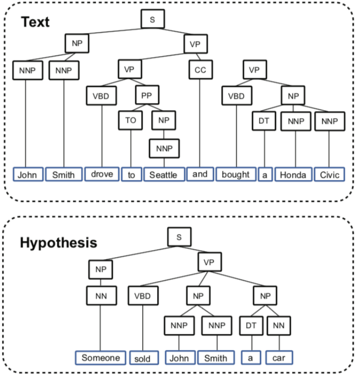
Lexical level

- H and T sentences encode aspects of underlying meaning that cannot be captured by the purely lexical representation.



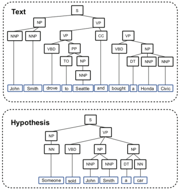
Structural level

- Syntactic structure provides cues for the underlying meaning of a sentence.



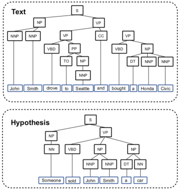
Structural level

- If T contains the same structure (i.e, dependency edges), the system will predict TRUE and otherwise FALSE.



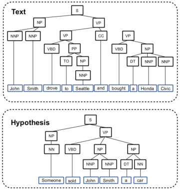
Structural level

- If T contains the same structure (i.e, dependency edges), the system will predict TRUE and otherwise FALSE.
- “John” and “drove,” but the two words are **separated** by a sequence of dependency edges.



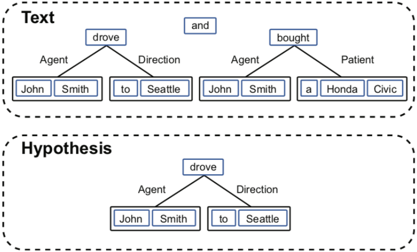
Structural level

- If T contains the same structure (i.e, dependency edges), the system will predict TRUE and otherwise FALSE.
- “John” and “drove,” but the two words are **separated** by a sequence of dependency edges.
- Given the expressiveness of the dependency representation, many possible sequences of edges that could represent connection, and many other sequences that do not.



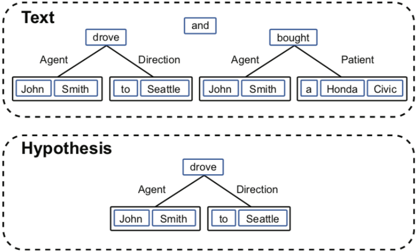
Semantic level

- Semantic role labelling, grouping of words into “arguments” (entity such as a person or place) and “predicates” (a predicate being a verb representing the state of some entity).



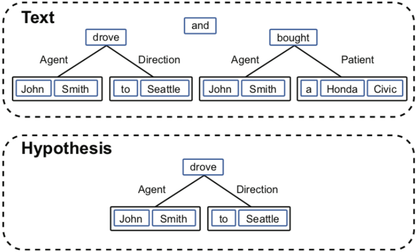
Semantic level

- Semantic role labelling, grouping of words into "arguments" (entity such as a person or place) and "predicates" (a predicate being a verb representing the state of some entity).
- **Immediate connections** between arguments and predicates.



Semantic level

- Semantic role labelling, grouping of words into "arguments" (entity such as a person or place) and "predicates" (a predicate being a verb representing the state of some entity).
- **Immediate connections** between arguments and predicates.
- "John" is an argument of the predicate "drove"



Knowledge Acquisition for RTE

- T: The U.S. citizens elected their new president Obama.
  H: Obama was born in the U.S.

Knowledge Acquisition for RTE

- T: The U.S. citizens elected their new president Obama.
  H: Obama was born in the U.S.
- Assumed **background knowledge**: "U.S. presidents should be naturally born in the U.S."

Knowledge Acquisition for RTE

- Knowledge is a lexical-semantic relation between two words.

Knowledge Acquisition for RTE

- Knowledge is a lexical-semantic relation between two words.
- I enlarged my **stock**. and I enlarged my **inventory**.
  synonym

Knowledge Acquisition for RTE

- Knowledge is a lexical-semantic relation between two words.
- I enlarged my **stock**. and I enlarged my **inventory**.
  synonym
- I have a **cat**. entails I have a **pet**.
  hyponymy

Knowledge Acquisition for RTE

- Knowledge is a lexical-semantic relation between two words.
- I enlarged my **stock**. and I enlarged my **inventory**.
  synonym
- I have a **cat**. entails I have a **pet**.
  hyponymy
- But also meaning implication between more complex structures than just lexical terms.
  X *causes* Y $\rightarrow$ Y *is a symptom of* X

Knowledge Acquisition for RTE

- WordNet specifies **lexical-semantic** relations between lexical items such as hyponymy, synonymy, and derivation.
  chair $\rightarrow$ furniture

Knowledge Acquisition for RTE

- WordNet specifies **lexical-semantic** relations between lexical items such as hyponymy, synonymy, and derivation.
  chair $\rightarrow$ furniture
- FrameNet is a lexicographic resource for **frames** that are events and includes information on the predicates and argument relevant for that specific event.
  The attack frame, and specifies events: 'assailant', a 'victim', a 'weapon', etc.
  cure X $\rightarrow$ X recover

Knowledge Acquisition for RTE

- WordNet specifies **lexical-semantic** relations between lexical items such as hyponymy, synonymy, and derivation.
  chair → furniture
- FrameNet is a lexicographic resource for **frames** that are events and includes information on the predicates and argument relevant for that specific event.
  The attack frame, and specifies events: 'assailant', a 'victim', a 'weapon', etc.
  cure X → X recover
- Wikipedia articles for identifying **is a** relations.
  Jim Carrey → actor

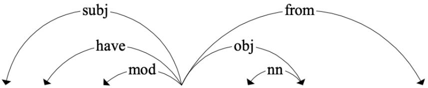
Knowledge Acquisition for RTE

- **Extended Distributional Hypothesis**: If two paths tend to occur in similar contexts, the meanings of the paths tend to be similar.

Knowledge Acquisition for RTE

- **Extended Distributional Hypothesis**: If two paths tend to occur in similar contexts, the meanings of the paths tend to be similar.
- X solves Y
  Y is solved by X
  X finds a solution to Y

Knowledge Acquisition for RTE



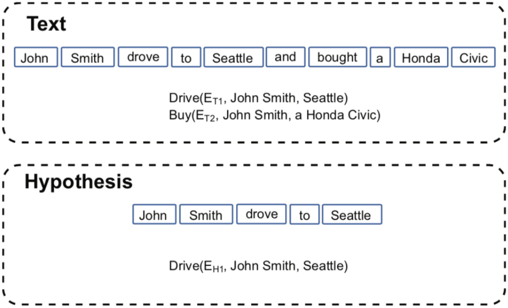
They had previously bought bighorn sheep from Comstock.

(a)	$\boxed{\text{N:subj:V}} \leftarrow \text{buy} \rightarrow \boxed{\text{V:from:N}}$	(X buys something from Y)
(b)	$\boxed{\text{N:subj:V}} \leftarrow \text{buy} \rightarrow \boxed{\text{V:obj:N}}$	(X buys Y)
(c)	$\boxed{\text{N:subj:V}} \leftarrow \text{buy} \rightarrow \boxed{\text{V:obj:N}} \rightarrow \text{sheep} \rightarrow \boxed{\text{N:nn:N}}$	(X buys Y sheep)
(d)	$\boxed{\text{N:nn:N}} \leftarrow \text{sheep} \leftarrow \boxed{\text{N:obj:V}} \leftarrow \text{buy} \rightarrow \boxed{\text{V:from:N}}$	(X sheep is bought from Y)
(e)	$\boxed{\text{N:obj:V}} \leftarrow \text{buy} \rightarrow \boxed{\text{V:from:N}}$	(X is bought from Y)

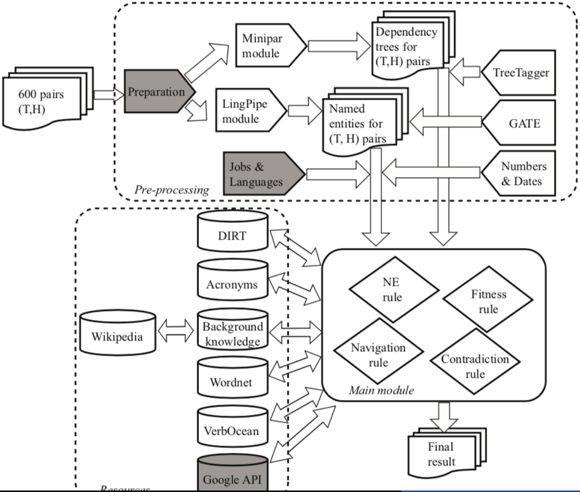
Outline

Recognising Textual Entailment Methods

- RTE depend on the representation (e.g. words, syntax, semantics) of the T-H pair that is used to extract features to train a supervised classifier.

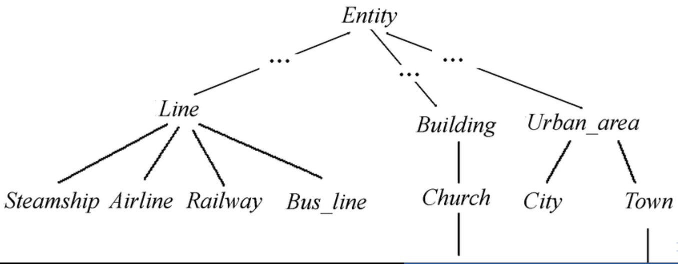


Recognising Textual Entailment Methods



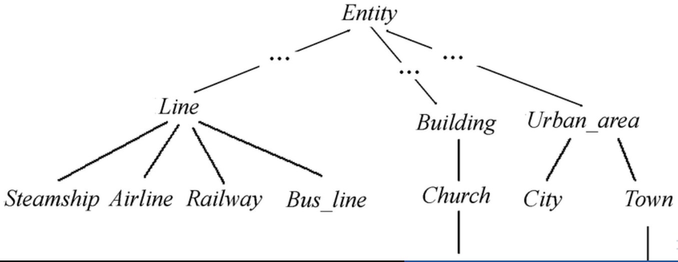
Similarity-based approaches

- Pair with a strong similarity score holds a positive entailment relation.



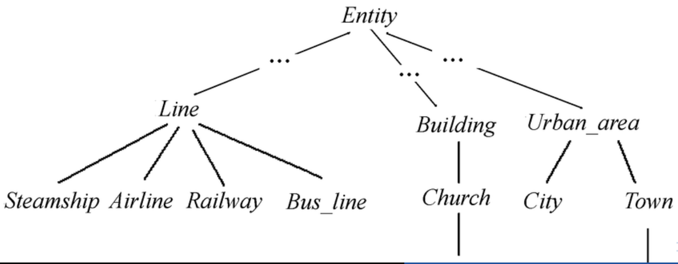
Similarity-based approaches

- Pair with a strong similarity score holds a positive entailment relation.
- Wordnet similarity.



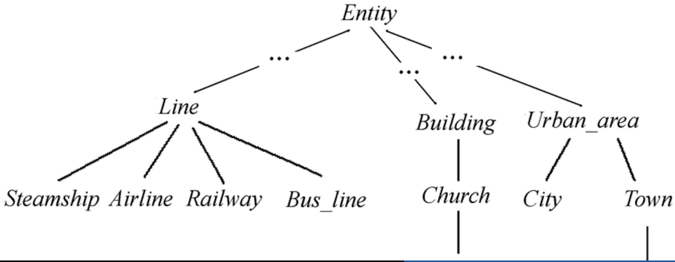
Similarity-based approaches

- Pair with a strong similarity score holds a positive entailment relation.
- Wordnet similarity.
- String similarity.

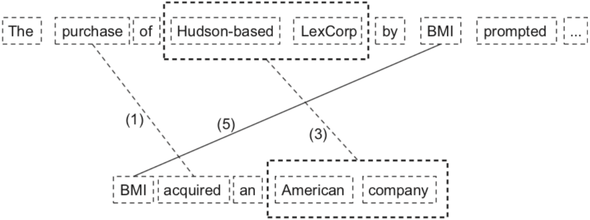


Similarity-based approaches

- Pair with a strong similarity score holds a positive entailment relation.
- Wordnet similarity.
- String similarity.
- Similarity scores computed from different linguistic levels. The goal is to find complementary features.



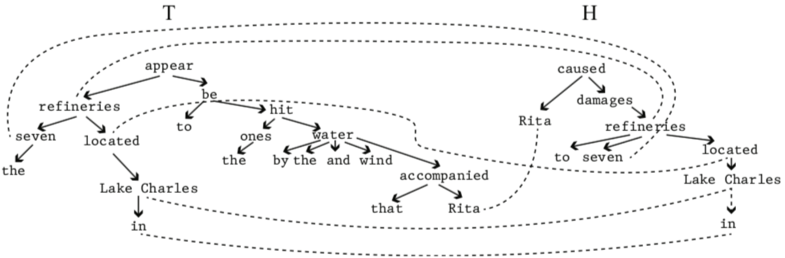
Alignment-based approaches



- (1,purchase,acquired)
  (3,Hudson-based LexCorp, American company),
  (5,BMI,BMI)

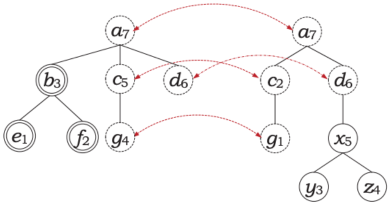
- $\rho_4 =$ purchase of $\boxed{\mathrm{X}}$ by $\boxed{\mathrm{Y}}$ $\rightarrow$ $\boxed{\mathrm{Y}}$ acquired $\boxed{\mathrm{X}}$
- $\rho_5 =$ $\boxed{\mathrm{Z{:}Noun}}$ of $\boxed{\mathrm{X}}$ by $\boxed{\mathrm{Y}}$ $\rightarrow$ $\boxed{\mathrm{Y}}$ $\boxed{\mathrm{Z{:}Verb}}$ $\boxed{\mathrm{X}}$

Alignment-based approaches



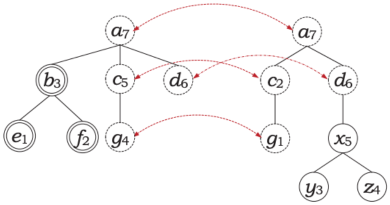
Edit distance-based approaches

- T entails H if there is a **sequence of transformations** applied to T such that we can obtain H with an overall cost below a certain **threshold**.



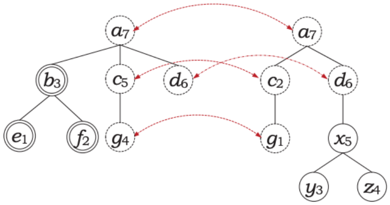
Edit distance-based approaches

- T entails H if there is a **sequence of transformations** applied to T such that we can obtain H with an overall cost below a certain **threshold**.
- Insertion, Substitution, and Deletion.



Edit distance-based approaches

- T entails H if there is a **sequence of transformations** applied to T such that we can obtain H with an overall cost below a certain **threshold**.
- Insertion, Substitution, and Deletion.
- Alternative for expensive theorem provers.



Evaluation

- Accuracy

Evaluation

- Accuracy
- RTE-3 corpus **1,600** T-H pairs
  information extraction, information retrieval, question answering, and summarisation.

Evaluation

- Accuracy
- RTE-3 corpus **1,600** T-H pairs
  information extraction, information retrieval, question answering, and summarisation.
- The lexical baseline, between 55% and 58% **accuracy**

Evaluation

- Accuracy
- RTE-3 corpus **1,600** T-H pairs
  information extraction, information retrieval, question answering, and summarisation.
- The lexical baseline, between 55% and 58% **accuracy**
- RTE-3 higher scores all system entries suggesting an **easier** entailment corpus

Evaluation

- Accuracy
- RTE-3 corpus **1,600** T-H pairs
  information extraction, information retrieval, question answering, and summarisation.
- The lexical baseline, between 55% and 58% **accuracy**
- RTE-3 higher scores all system entries suggesting an **easier** entailment corpus
- RTE-4 and RTE-5 increase the difficulty by adding irrelevant signals (additional words, phrases, and sentences).

Outline

SNLI

- Flickr30k corpus for image captioning domaim. Annotated pairs of texts at sentence level

SNLI

- Flickr30k corpus for image captioning domaim.
  Annotated pairs of texts at sentence level
- The relations (i.e. 3-way classification labels) are:
  entailment, *contradiction*, and *neutral*.

SNLI

- Flickr30k corpus for image captioning domaim. Annotated pairs of texts at sentence level
- The relations (i.e. 3-way classification labels) are: *entailment*, *contradiction*, and *neutral*.
- $550,152$ training, 10K development, and 10k test.

SNLI

- Flickr30k corpus for image captioning domaim. Annotated pairs of texts at sentence level
- The relations (i.e. 3-way classification labels) are: *entailment*, *contradiction*, and *neutral*.
- $550,152$ training, 10K development, and 10k test.
- Premise: A soccer game with multiple males playing. Hypothesis: Some men are playing a sport.

MNLI

- Multiple genres
  classifiers only learn **regularities** over annotated data, leading to **poor generalization** beyond the domain of the training data

MNLI

- Multiple genres
  classifiers only learn **regularities** over annotated data, leading to **poor generalization** beyond the domain of the training data
- *matched* (5 in domain genres) $392,702$ training,
  10k *matched* development,

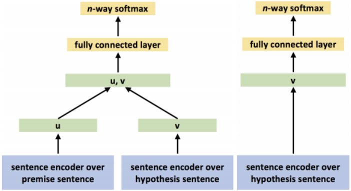
MNLI

- Multiple genres
  classifiers only learn **regularities** over annotated data, leading to **poor generalization** beyond the domain of the training data
- *matched* (5 in domain genres) $392,702$ training,
  10k *matched* development,
- 10k *mismatched* (5 out of domain genres) development.

MNLI

- Multiple genres
  classifiers only learn **regularities** over annotated data, leading to **poor generalization** beyond the domain of the training data
- *matched* (5 in domain genres) 392,702 training,
  10k *matched* development,
- 10k *mismatched* (5 out of domain genres) development.
- T: 8 million in relief in the form of emergency housing.
  H: The 8 million dollars for emergency housing was still not enough to solve the problem.
  Government

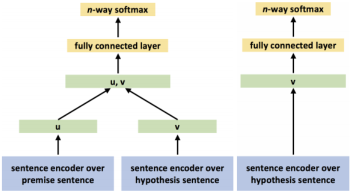
Drawbacks



Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Entailment	There are **at least** three **people** on a loading dock.
Neutral	A woman is selling bamboo sticks **to help provide for her family.**
Contradiction	A woman is **not** taking money for any of her sticks.

- Entailment: animal, instrument, and outdoors.

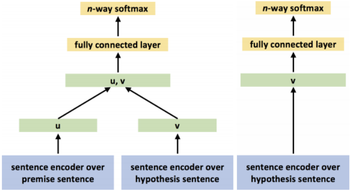
Drawbacks



Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Entailment	There are **at least** three **people** on a loading dock.
Neutral	A woman is selling bamboo sticks **to help provide for her family.**
Contradiction	A woman is **not** taking money for any of her sticks.

- Entailment: animal, instrument, and outdoors.
- Neutral: Modifiers (tall, sad, popular) and superlatives (first, favorite, most)

Drawbacks

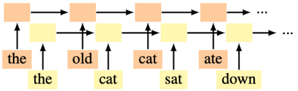


Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Entailment	There are **at least** three **people** on a loading dock.
Neutral	A woman is selling bamboo sticks **to help provide for her family.**
Contradiction	A woman is **not** taking money for any of her sticks.

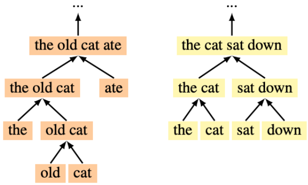
- Entailment: animal, instrument, and outdoors.
- Neutral: Modifiers (tall, sad, popular) and superlatives (first, favorite, most)
- Contradiction: Negation words, nobody, no, never and nothing

Neural Network Models

- Embeddings like glove or elmo, for fine tuning.



(a) A conventional sequence-based RNN for two sentences.

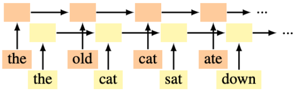


]

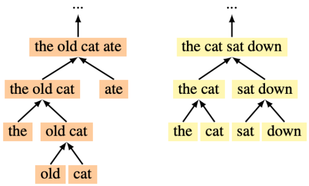
(b) A conventional TreeRNN for two sentences.

Neural Network Models

- Embeddings like glove or elmo, for fine tuning.
- Sentence representations.



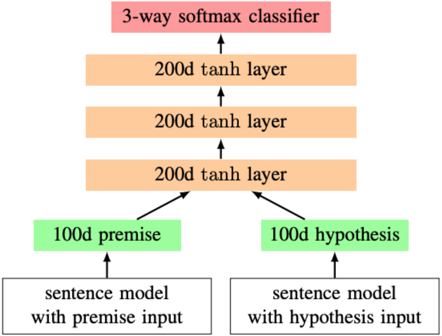
(a) A conventional sequence-based RNN for two sentences.

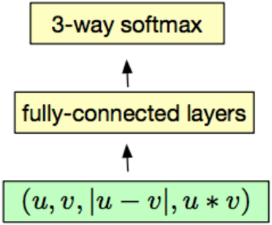


]

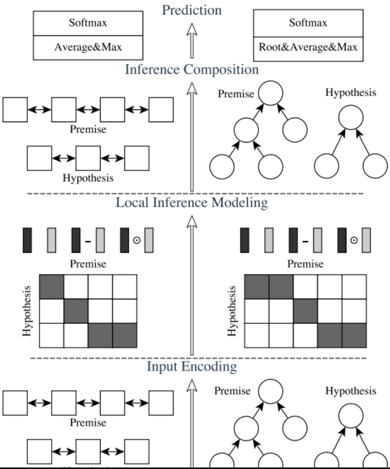
(b) A conventional TreeRNN for two sentences.

BiLSMT composition





ESIM

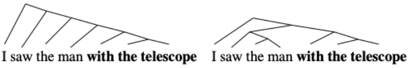

Softmax
Average&Max
Prediction
Softmax
Root&Average&Max
Inference Composition
Premise
Hypothesis
Premise
Hypothesis
Local Inference Modeling
Premise
Hypothesis
Premise
Hypothesis
Input Encoding
Premise
Hypothesis
Premise

ESIM

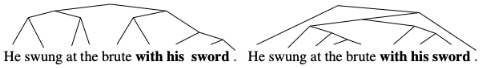
$$
\begin{aligned}
\mathbf{t}_i &= \mathrm{emb}(t_i; \omega_{\mathrm{emb}}) && \text{(1a)}\\
\mathbf{h}_j &= \mathrm{emb}(h_j; \omega_{\mathrm{emb}}) && \text{(1b)}\\
\mathbf{s}_1^m &= \mathrm{birnn}(\mathbf{t}_1^m; \omega_{\mathrm{enc}}) && \text{(1c)}\\
\mathbf{u}_1^n &= \mathrm{birnn}(\mathbf{h}_1^n; \omega_{\mathrm{enc}}) && \text{(1d)}\\
\mathbf{a}_i &= \mathrm{attention}(\mathbf{s}_i, \mathbf{u}_1^n) && \text{(1e)}\\
\mathbf{b}_j &= \mathrm{attention}(\mathbf{u}_j, \mathbf{s}_1^m) && \text{(1f)}\\
\mathbf{c}_i &= [\mathbf{s}_i, \mathbf{a}_i, \mathbf{s}_i - \mathbf{a}_i, \mathbf{s}_i \odot \mathbf{a}_i] && \text{(1g)}\\
\mathbf{d}_j &= [\mathbf{u}_j, \mathbf{b}_j, \mathbf{u}_j - \mathbf{b}_j, \mathbf{u}_j \odot \mathbf{b}_j] && \text{(1h)}\\
\mathbf{c}_1^m &= \mathrm{birnn}(\mathbf{c}_1^m; \omega_{\mathrm{comp}}) && \text{(1i)}\\
\mathbf{d}_1^n &= \mathrm{birnn}(\mathbf{d}_1^n; \omega_{\mathrm{comp}}) && \text{(1j)}\\
\mathbf{q} &= [\mathrm{avg}(\mathbf{c}_1^m), \mathrm{maxpool}(\mathbf{c}_1^m), \mathrm{avg}(\mathbf{d}_1^n), \mathrm{maxpool}(\mathbf{d}_1^n)] && \text{(1k)}\\
\mathbf{q} &= \tanh(\mathrm{affine}(\mathbf{q}; \omega_{\mathrm{hid}})) && \text{(1l)}\\
f(x) &= \mathrm{softmax}(\mathrm{mlp}(\mathbf{q}; \omega_{\mathrm{cls}})) && \text{(1m)}
\end{aligned}
$$

Outline

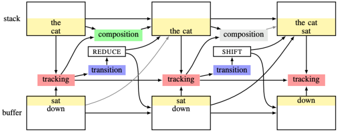
Latent Structure Induction

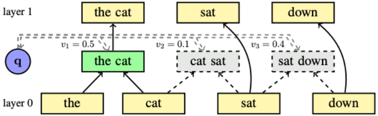


(a) Two parse trees correspond to two distinct interpretations for the sentence in example (1).



(b) Parses generated by at ST-Gumbel model (left) and the Stanford Parser (right).





Deep Generative Models

- Model that generates **hypothesis** and **decision** given a text and a stochastic embedding of the hypothesis-decision pair.

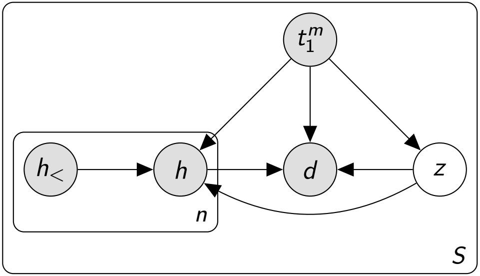
Deep Generative Models

- Model that generates **hypothesis** and **decision** given a text and a stochastic embedding of the hypothesis-decision pair.
- Models to learn from mixed-domain NLI data
  e.g. by capitalising on lexical domain-dependent patterns.

Deep Generative Models

- Model that generates **hypothesis** and **decision** given a text and a stochastic embedding of the hypothesis-decision pair.
- Models to learn from mixed-domain NLI data
  e.g. by capitalising on lexical domain-dependent patterns.
- Performance of standard classifiers tend to vary across domains and especially out of domain.

Deep Generative Models



$$Z_i | t_1^m \sim \mathcal{N}(\mu(s_1^m), \sigma^2(s_1^m))$$
$$H_i | z_1^m \sim Cat(f(z_1^m, t_1^m; \theta))$$
$$D_j | z_1^m, h_1^n \sim Cat(g(z_1^m, t_1^m, h_1^n; \theta))$$

Deep Generative Models I

- Joint likelihood of y (hypothesis) and d (decision)

$$p(y, d|x, \theta) = \int p(z|x, \theta) p(y|x, z, \theta) p(d|x, y, z, \theta) dz. \tag{2}$$

- The *hypothesis generation model*:

$$p(y|x, z, \theta) = \prod_{j=1}^{|y|} p(y_j|x, z, y_{<j}, \theta) = \prod_{j=1}^{|y|} \mathrm{Cat}(y_j|f_o(x, z, y_{<j}; \theta)) \ , \tag{3}$$

Deep Generative Models II

- The *classification model* ESIM:

$$p(d|x, y, z, \theta) = \mathrm{Cat}(d|f_c(x, y, z; \theta)) \tag{4}$$

- Lowerbound on the log-likelihood function (ELBO)

$$\begin{aligned}\mathcal{L}(\theta, \phi) = {} & \mathbb{E}_{q(z|x,y,d,\phi)}\left[\log p(y, d|x, z, \theta)\right] \\ & - \mathrm{KL}(q(z|x, y, d, \phi)||p(z|x, \theta))\end{aligned} \tag{5}$$

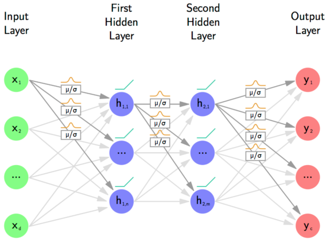
Deep Generative Models

Model	Dev	
	matched	mismatched
$\text{ESIM}_{\text{mnli}}$	74.39 ± 0.11	74.05 ± 0.21
$+\ \mathcal{N}\text{-VAE}_{50z}$	74.89 ± 0.25	74.07 ± 0.37
$+\ \mathcal{N}\text{-VAE}_{100z}$	74.82 ± 0.28	73.91 ± 0.59
$+\ \mathcal{N}\text{-VAE}_{256z}$	74.87 ± 0.15	74.08 ± 0.16

Outline

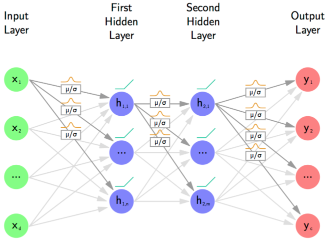
Bayes by backprop

- NNs perform well with lots of data, however they fail to express uncertainty with little or no data,
  leading to overconfident decisions.



Bayes by backprop

- NNs perform well with lots of data, however they fail to express uncertainty with little or no data, leading to overconfident decisions.
- Bayesian neural networks introduce probability distributions over the weights.



Bayes by backprop

- However, Bayesian inference on the parameters ω of a neural network is intractable, with data D.

$$p(\omega|\mathcal{D}) = \frac{p(\mathcal{D}|\omega)p(\omega)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\omega)p(\omega)}{\int p(\mathcal{D}|\omega)p(\omega)\mathrm{d}\omega} \quad (6)$$

(Blundell et al., 2015)

Bayes by backprop

- However, Bayesian inference on the parameters ω of a neural network is intractable, with data D.

$$p(\omega|\mathcal{D}) = \frac{p(\mathcal{D}|\omega)p(\omega)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\omega)p(\omega)}{\int p(\mathcal{D}|\omega)p(\omega)\mathrm{d}\omega} \tag{6}$$

- We need an approximation $q(\omega|\theta)$, over the weights that approximates the true posterior

(Blundell et al., 2015)

Bayes by backprop

- However, Bayesian inference on the parameters ω of a neural network is intractable, with data D.

$$p(\omega|\mathcal{D}) = \frac{p(\mathcal{D}|\omega)p(\omega)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\omega)p(\omega)}{\int p(\mathcal{D}|\omega)p(\omega)\mathrm{d}\omega} \tag{6}$$

- We need an approximation $q(\omega|\theta)$, over the weights that approximates the true posterior
- The ELBO is:

$$\begin{aligned}\mathcal{L}(\mathcal{D},\theta) &= \int q(\omega|\theta)\log\frac{q(\omega|\theta)}{p(\omega)} - q(\omega|\theta)\log p(\mathcal{D}|\omega)\mathrm{d}\omega \\ &= \mathrm{KL}[q(\omega|\theta)||p(\omega)] - \mathbb{E}_{q(\omega|\theta)}[\log p(\mathcal{D}|\omega)]\end{aligned} \tag{7}$$

(Blundell et al., 2015)

MC dropout I

- On NLI training inputs $X = \langle (t_1, h_1), \ldots, (t_n, h_n) \rangle$ are premise (t) and hypothesis (h) pairs, and the corresponding outputs $Y = \langle y_1, \ldots, y_n \rangle$ over N instances.
- The likelihood for classification is defined by:

$$p(y|x, \omega) = \mathrm{Cat}(y|f(x; \omega)), \tag{8}$$

over y entailment relations computed by mapping from the input to the class probabilities with a neural network f parameterised by ω.

MC dropout II

- A Bayesian NN (MacKay, 1992) is defined by placing a prior distribution over the model parameters $p(\omega)$, where this prior is often a Gaussian distribution $p(\omega) \sim \mathcal{N}(0, I)$.
- The Bayesian NN formulation leads to a posterior distribution over the parameters given our observed data, instead of a single estimate.
- We are interested on estimating the posterior distribution over the parameters $p(\omega|\mathcal{D})$, given our observed data X, Y.
- The goal is to predict a new input instances by marginalising over the parameters:

$$p(y^*|x^*, \mathcal{D}) = \int p(y^*|x^*, \omega) p(\omega|\mathcal{D}) d\omega. \tag{9}$$

MC dropout III

- However, the true posterior $p(\omega|\mathcal{D})$ is intractable, and Gal and Ghahramani (2016a) use variational inference to approximate this posterior.
- We define an approximate distribution $q_\theta(\omega)$, to minimise the KL divergence between the approximation and the true posterior.
- The objective for optimisation is a lower-bound on the log-likelihood function (ELBO):

$$\mathcal{L} = \mathbb{E}_{q(\omega)}\left[\sum_{i=1}^{N} \log p(y_i|f(x_i;\omega))\right] - \mathrm{KL}(q_\theta(\omega))||p(\omega)), \tag{10}$$

MC dropout IV

where the KL term is approximated with L_2 regularisation.

- Gal and Ghahramani (2016a) show that the use of dropout in NNs before each weight layer is an approximation to variational inference in Bayesian NNs.
- By replacing the true posterior $p(\omega|\mathcal{D})$ with the approximate posterior $q_\theta(\omega)$, we obtain a Monte Carlo (MC) estimate for future predictions :

$$\begin{aligned} p(y^*|x^*, \mathcal{D}) &\approx \int p(y^*|x^*, \omega) q_\theta(\omega) d\omega \\ &\approx \frac{1}{T} \sum_t^T p(y^*|x^*, \hat{\omega}_t), \end{aligned} \tag{11}$$

MC dropout V

where $\hat{\omega}_t \sim q_\theta(\omega)$

- In practice, the approximation to the predictive distribution is based on performing T stochastic forward passes through the network and averaging the results.
- In other words, this is achieved by performing **dropout at test time** (MC dropout).
- Finally, for classification, a way to quantify uncertainty is by computing the entropy of the output probability vector $\mathcal{H}(p) = -\sum_{c=1}^{C} p_c \log p_c$ over c classes.

Uncertainty in natural language inference

- ESIM for classification (without syntactic parses)

Uncertainty in natural language inference

- ESIM for classification (without syntactic parses)
- The word embedding and the bidirectional LSTMs are shared between the pair of texts. A single (tanh) hidden layer MLP with a softmax output predicts the class probabilities.

Uncertainty in natural language inference

- ESIM for classification (without syntactic parses)
- The word embedding and the bidirectional LSTMs are shared between the pair of texts. A single (tanh) hidden layer MLP with a softmax output predicts the class probabilities.
- We use dropout on both the LSTM (variational RNN), and the word embedding.

Uncertainty in natural language inference

- ESIM for classification (without syntactic parses)
- The word embedding and the bidirectional LSTMs are shared between the pair of texts. A single (tanh) hidden layer MLP with a softmax output predicts the class probabilities.
- We use dropout on both the LSTM (variational RNN), and the word embedding.
- In the word embedding $\omega_{\text{emb}} \in \mathrm{R}^{V \times D}$, with V vocabulary and D dimensionality, the dropout masks types (rows) instead of words in a sequence .

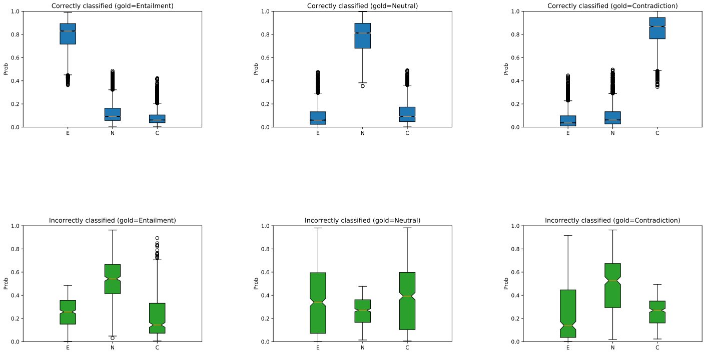
Uncertainty in natural language inference

- ESIM for classification (without syntactic parses)
- The word embedding and the bidirectional LSTMs are shared between the pair of texts. A single (tanh) hidden layer MLP with a softmax output predicts the class probabilities.
- We use dropout on both the LSTM (variational RNN), and the word embedding.
- In the word embedding $\omega_{\text{emb}} \in \mathrm{R}^{V \times D}$, with V vocabulary and D dimensionality, the dropout masks types (rows) instead of words in a sequence .
- Finally, for the additional L_2 regularisation, we use a separate weight decay: for weights $\lambda_\omega = \frac{1-p_{\text{drop}}}{N}$ with p_{drop} dropout, and for biases (b): $\lambda_{\text{b}} = \frac{1}{N}$.

Results

Training	Model	SNLI	Breaking NLI
SNLI	ESIM†	87.9	65.6
	$\text{ESIM}_{\text{ours}}$	86.4 ± 0.09	57.6 ± 1.9
	ESIM_{MC}	86.5 ± 0.13	68.9 ± 1.7
MNLI+SNLI	ESIM†	86.3	74.9
	$\text{ESIM}_{\text{ours}}$	86.8 ± 0.05	68.8 ± 3.5
	ESIM_{MC}	86.6 ± 0.16	75.2 ± 1.3

Results SNLI

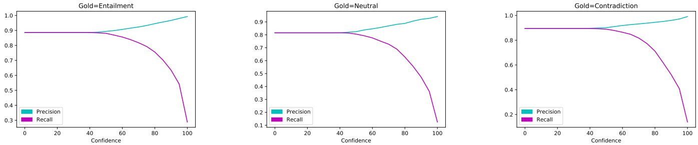


(d) Gold label entailment.

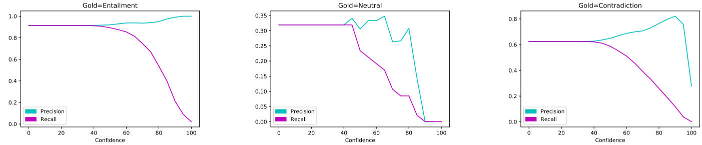
(e) Gold label neutral.

(f) Gold label contradiction.

Results SNLI and Breaking



(g) SNLI



(j) Breaking

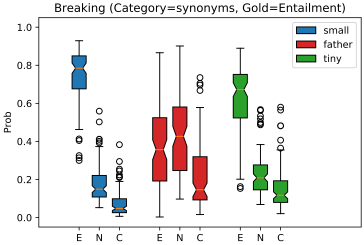
Results

P: The little girl is riding in the car with her dad.
H: The small girl is riding in the car with her dad.

P: The little girl is riding in the car with her dad.
H: The little girl is riding in the car with her father.

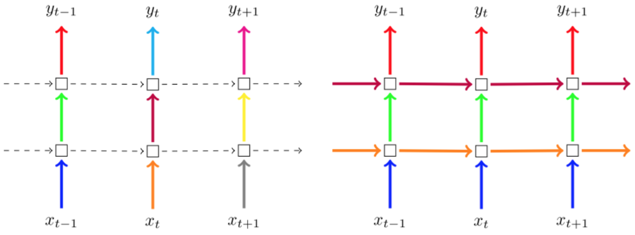
P: The little girl is riding in the car with her dad.

H: The tiny girl is riding in the car with her dad.



Homework!!

- Dropout in Recurrent Networks (Gal and Ghahramani, 2016b)
- Use the same dropout mask at each time step for both inputs, outputs, and recurrent layers
- The RNN can be framed as a probabilistic model.



(a) Naive dropout RNN (b) Variational RNN

Literature I

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 1613–1622. JMLR.org, 2015. URL `http://dl.acm.org/citation.cfm?id=3045118.3045290`.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 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, 2017.

Literature II

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016a. PMLR. URL `http://proceedings.mlr.press/v48/gal16.html`.

Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in neural information processing systems*, pages 1019–1027, 2016b.

Literature III

Max Glockner, Vered Shwartz, and Yoav Goldberg. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia, July 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P18-2103.

David J. C. MacKay. A practical bayesian framework for backpropagation networks. *Neural Comput.*, 4(3):448–472, May 1992. ISSN 0899-7667. doi: 10.1162/neco.1992.4.3.448. URL http://dx.doi.org/10.1162/neco.1992.4.3.448.