Textual Entailment and Paraphrasing

NLP1

Miguel Rios
m.riosgaona@uva.nl
University of Amsterdam

November 29, 2018

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

1. Introduction
   - Applications of Textual Entailment
2. Levels of Representation
3. RTE Methods
   - Evaluation
4. Current Methods
   - NLI Datasets
   - Neural Network Models
   - Latent Variable Models
   - Drawbacks

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

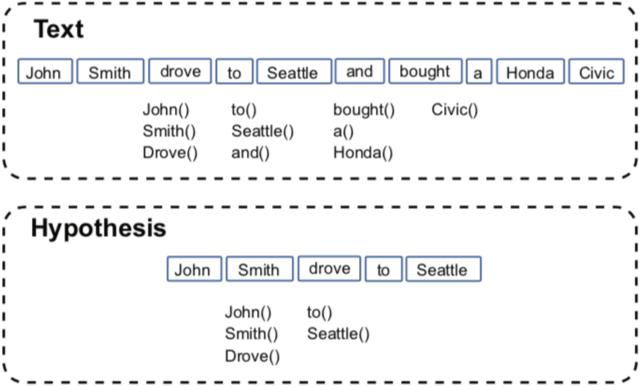
- Determining the equivalence or non-equivalence of the meanings of the T-H.

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.

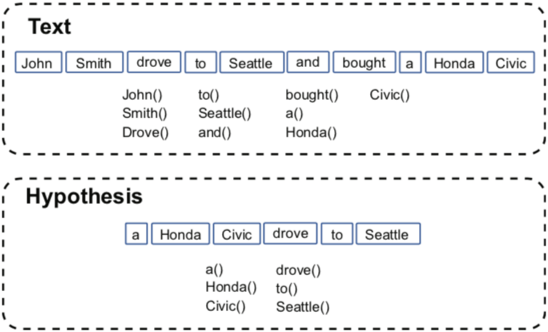
Lexical level

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



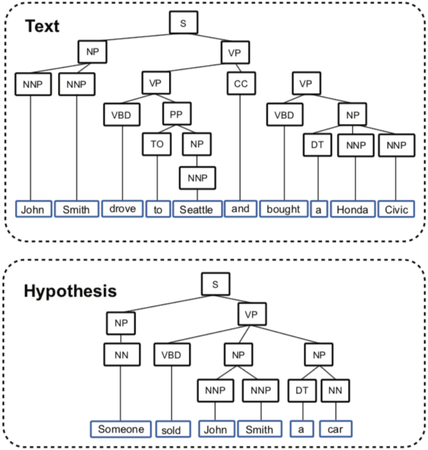
Lexical level

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



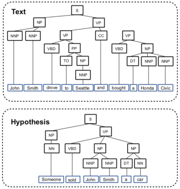
Structural level

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



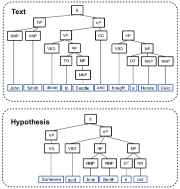
Structural level

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



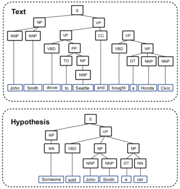
Structural level

- 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.



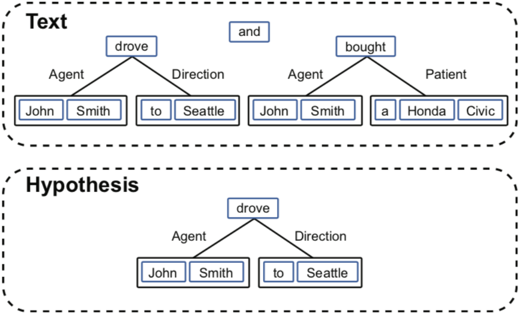
Structural level

- 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.



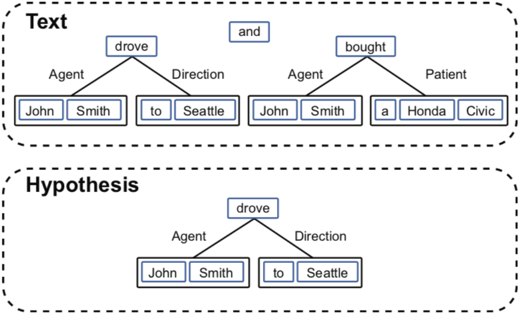
Semantic level

- 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).



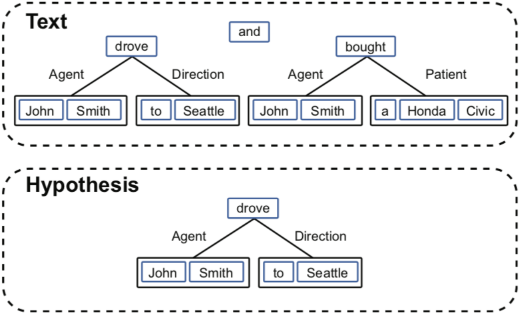
Semantic level

- 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

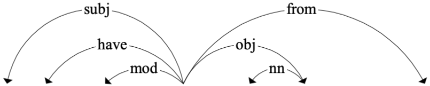
Knowledge Acquisition for RTE

- **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



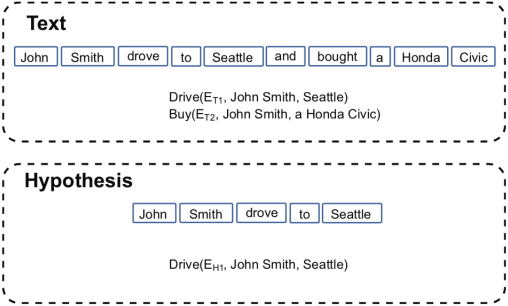
They had previously bought bighorn sheep from Comstock.

(a)	N:subj:V←buy→V:from:N	(*X* buys something from *Y*)
(b)	N:subj:V←buy→V:obj:N	(*X* buys *Y*)
(c)	N:subj:V←buy→V:obj:N→sheep→N:nn:N	(*X* buys *Y* sheep)
(d)	N:nn:N←sheep←N:obj:V←buy→V:from:N	(*X* sheep is bought from *Y*)
(e)	N:obj:V←buy→V:from:N	(*X* is bought from *Y*)

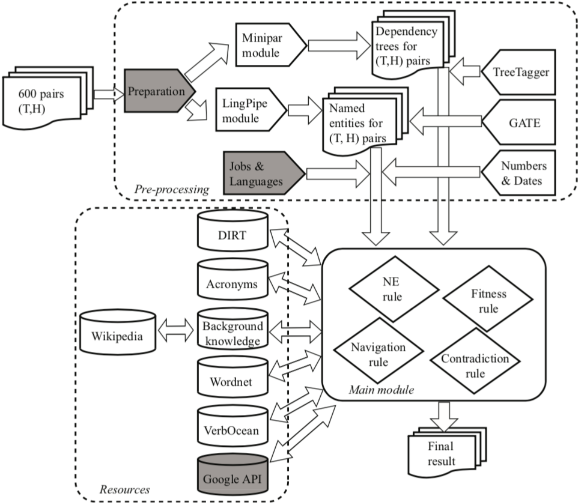
Outline

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.

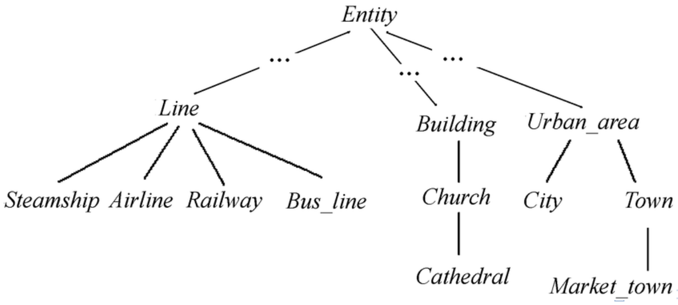


Recognising Textual Entailment Methods



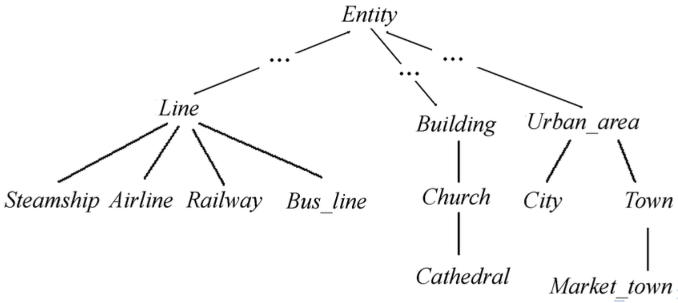
Similarity-based approaches

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



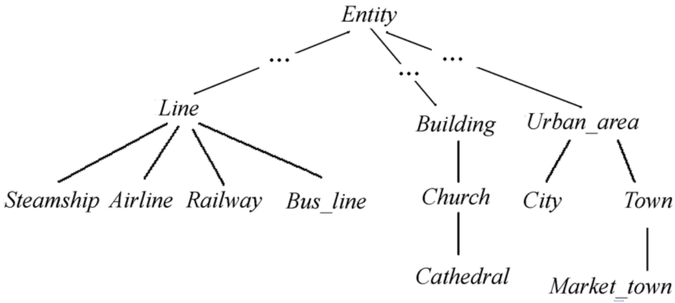
Similarity-based approaches

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



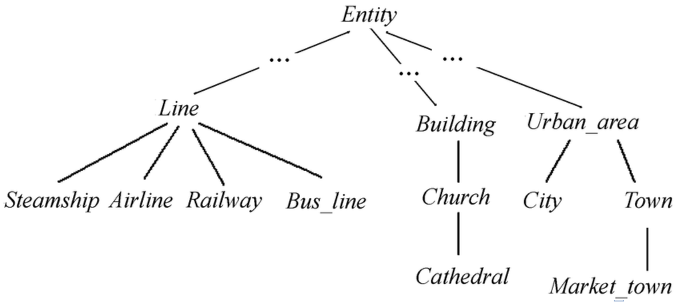
Similarity-based approaches

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

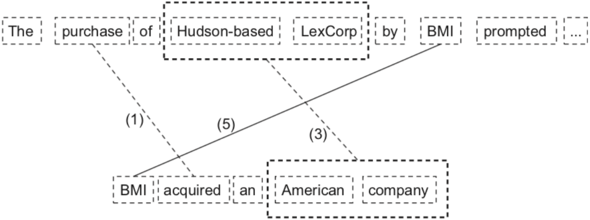


Similarity-based approaches

- 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.



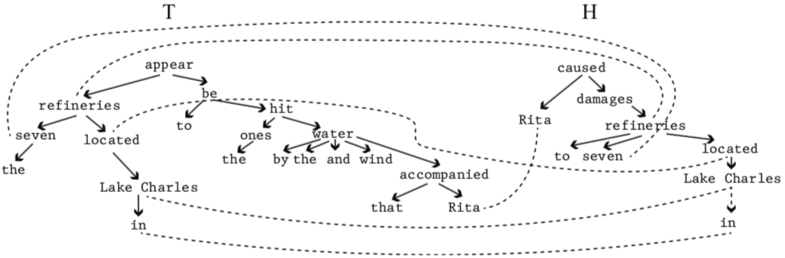
Alignment-based approaches



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

- ρ_4 = purchase of $\boxed{X}$ by $\boxed{Y}$ $\rightarrow$ $\boxed{Y}$ acquired $\boxed{X}$
- ρ_5 = $\boxed{Z{:}Noun}$ of $\boxed{X}$ by $\boxed{Y}$ $\rightarrow$ $\boxed{Y}$ $\boxed{Z{:}Verb}$ $\boxed{X}$

Alignment-based approaches



Formal Logic approaches

- Search for whether or not the entailment holds by finding proofs with a theorem prover.

Text: John bought a Jeep.

Hyp: John owns a car.

World Knowledge:
1. ∀ X Jeep(X) → car(X)
(a Jeep is a car)
2. ∀ X,Y buy(X,Y) → own(X,Y)
(if X buys Y, then X owns Y)

Proof: Initial state of belief (Text):
∃ A,B John(A) ∧ Jeep(B) ∧ buy(A,B)
(John bought a Jeep)

Target assertion:
∃ C,D John(C) ∧ car(D) ∧ own(C,D)

1. apply rule 1 with Text clause "Jeep(B)":
Jeep(B) → car(B)

State of belief is now:
∃ A,B John(A) ∧ car(B) ∧ buy(A,B)

2. apply rule 2 with Text clause "buy(A,B)":
buy(A,B) → own(A,B)

State of belief is now:
∃ A,B John(A) ∧ car(B) ∧ own(A,B)

Formal Logic approaches

- Search for whether or not the entailment holds by finding proofs with a theorem prover.
- **Lack** of background knowledge (recall vs. precision) which not many true decisions could be found.

Text: John bought a Jeep.

Hyp: John owns a car.

World Knowledge:
1. ∀ X Jeep(X) → car(X)
(a Jeep is a car)
2. ∀ X,Y buy(X,Y) → own(X,Y)
(if X buys Y, then X owns Y)

Proof: Initial state of belief (Text):
∃ A,B John(A) ∧ Jeep(B) ∧ buy(A,B)
(John bought a Jeep)

Target assertion:
∃ C,D John(C) ∧ car(D) ∧ own(C,D)

1. apply rule 1 with Text clause "Jeep(B)":
Jeep(B) → car(B)

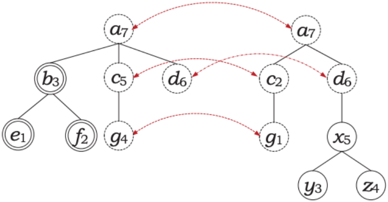
State of belief is now:
∃ A,B John(A) ∧ car(B) ∧ buy(A,B)

2. apply rule 2 with Text clause "buy(A,B)":
buy(A,B) → own(A,B)

State of belief is now:
∃ A,B John(A) ∧ car(B) ∧ own(A,B)

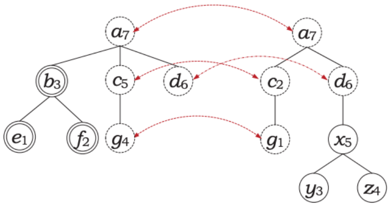
Edit distance-based approaches

- 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**.



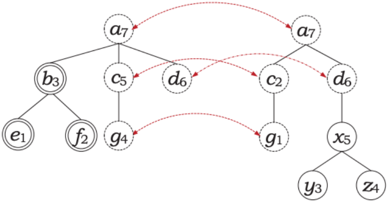
Edit distance-based approaches

- 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

Current Methods

TL;DR

- Limitation with small data in RTE.

Current Methods

TL;DR

- Limitation with small data in RTE.
- Availability of large **NLI** data.

Current Methods

TL;DR

- Limitation with small data in RTE.
- Availability of large **NLI** data.
- **Flexibility** of Neural Network models.

Current Methods

TL;DR

- Limitation with small data in RTE.
- Availability of large **NLI** data.
- **Flexibility** of Neural Network models.
- **Lack** of knowledge in NN models .

Current Methods

TL;DR

- Limitation with small data in RTE.
- Availability of large **NLI** data.
- **Flexibility** of Neural Network models.
- **Lack** of knowledge in NN models .
- **Induce structure** or representation given NLI task.

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

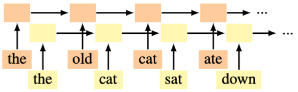
- Multiple genres
  classifiers only learn **regularities** over annotated data, leading to **poor generalization** beyond the domain of the training data
- *matched* (5 in domain genres) $392,702$ training,
  10k *matched* development,
- 10k *mismatched* (5 out of domain genres) development.

MNLI

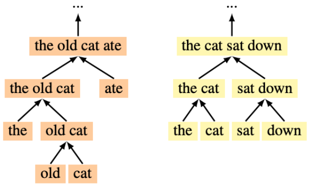
- Multiple genres
  classifiers only learn **regularities** over annotated data, leading to **poor generalization** beyond the domain of the training data
- *matched* (5 in domain genres) $392,702$ training,
  10k *matched* development,
- 10k *mismatched* (5 out of domain genres) development.
- T: 8 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

Neural Network Models

- Embeddings like glove or elmo, for fine tuning.



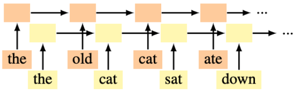
(a) A conventional sequence-based RNN for two sentences.



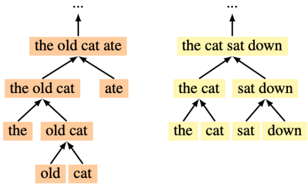
]
(b) A conventional TreeRNN for two sentences.

Neural Network Models

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



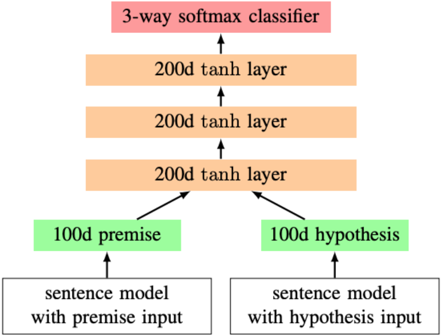
(a) A conventional sequence-based RNN for two sentences.

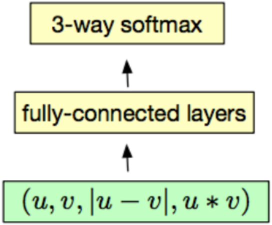


]

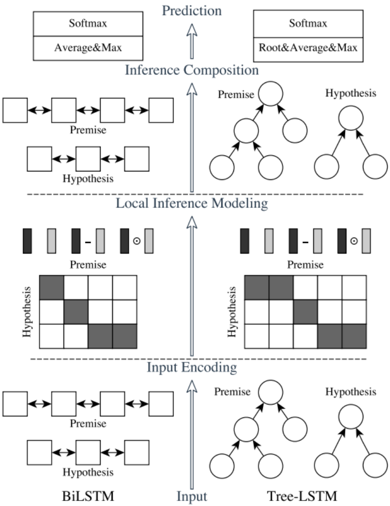
(b) A conventional TreeRNN for two sentences.

BiLSMT composition

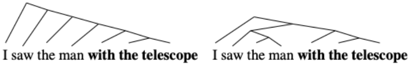




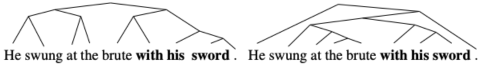
ESIM


Prediction
Softmax
Average&Max
Softmax
Root&Average&Max
Inference Composition
Premise
Hypothesis
Premise
Hypothesis
Local Inference Modeling
Premise
Hypothesis
Premise
Hypothesis
Input Encoding
Premise
Hypothesis
Premise
Hypothesis
BiLSTM
Input
Tree-LSTM

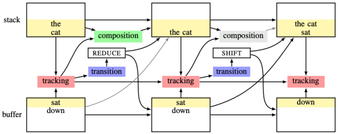
Latent Structure Induction

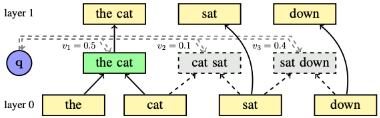


(a) Two parse trees correspond to two distinct interpretations for the sentence in example (1).



(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.

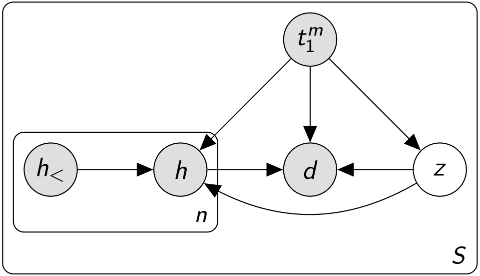
Deep Generative Models

- 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

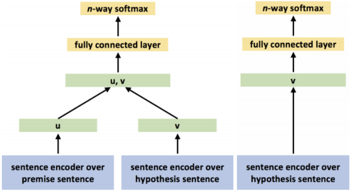


$$Z_i | t_1^m \sim \mathcal{N}(\mu(s_1^m), \sigma^2(s_1^m))$$
$$H_i | z_1^m \sim Cat(f(z_1^m, t_1^m; \theta))$$
$$D_j | z_1^m, h_1^n \sim Cat(g(z_1^m, t_1^m, h_1^n; \theta))$$

Deep Generative Models

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
- disentanglement	74.47 ± 0.13	73.88 ± 0.55
$+\ \mathcal{S}\text{-VAE}_{10z,1.0k}$	74.82 ± 0.11	74.59 ± 0.02

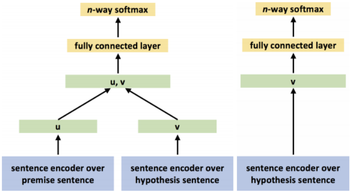
Drawbacks



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.

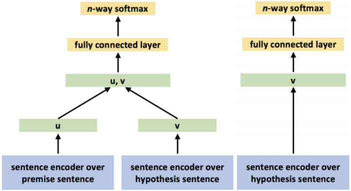
Drawbacks



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