Two Aspects of Text Representations for NLP and MT: Morphology and Deep Learning

Hinrich Schütze

Center for Information and Language Processing, LMU Munich

2015-09-10

Text Representations for NLP and MT

- How should the input to NLP and MT systems be represented?

Text Representations for NLP and MT

- How should the input to NLP and MT systems be represented?
- Statistical NLP/MT

Text Representations for NLP and MT

- How should the input to NLP and MT systems be represented?
- Statistical NLP/MT
  - The representation should make generalization easy.

Text Representations for NLP and MT

- How should the input to NLP and MT systems be represented?
- Statistical NLP/MT
  - The representation should make generalization easy.
- Rule-based NLP/MT

Text Representations for NLP and MT

- How should the input to NLP and MT systems be represented?
- Statistical NLP/MT
  - The representation should make generalization easy.
- Rule-based NLP/MT
  - The representation should make it easy to formulate accurate, broad-coverage rules/constraints.

Text Representations for NLP and MT

- How should the input to NLP and MT systems be represented?
- Statistical NLP/MT
  - The representation should make generalization easy.
- Rule-based NLP/MT
  - The representation should make it easy to formulate accurate, broad-coverage rules/constraints.
- Topic of this talk: two aspects of “good” representation

Text Representations for NLP and MT

- How should the input to NLP and MT systems be represented?
- Statistical NLP/MT
  - The representation should make generalization easy.
- Rule-based NLP/MT
  - The representation should make it easy to formulate accurate, broad-coverage rules/constraints.
- Topic of this talk: two aspects of “good” representation
  - morphology

Text Representations for NLP and MT

- How should the input to NLP and MT systems be represented?
- Statistical NLP/MT
  - The representation should make generalization easy.
- Rule-based NLP/MT
  - The representation should make it easy to formulate accurate, broad-coverage rules/constraints.
- Topic of this talk: two aspects of "good" representation
  - morphology
  - deep learning embeddings

Overview

1. Morphology
2. Deep learning embeddings
3. Morphological lexica vs embeddings
4. For units of which granularities should we use embeddings?
5. Using deep learning (in general) in MT

Disclaimer

I am not an MT researcher!

Outline

1. Morphology
2. Deep learning embeddings
3. Morphological lexica vs embeddings
4. For units of which granularities should we use embeddings?
5. Using deep learning (in general) in MT

Why worry about morphology in MT

Why worry about morphology in MT

- Much of statistical NLP:
  estimate and use $p_\theta(y|x)$, $y \in Y$, $x \in X$

Why worry about morphology in MT

- Much of statistical NLP:
  estimate and use $p_\theta(y|x)$, $y \in Y$, $x \in X$
- X = representation for language, including words

Why worry about morphology in MT

- Much of statistical NLP:
  estimate and use $p_\theta(y|x)$, $y \in Y$, $x \in X$
- X = representation for language, including words
- Y = some event / fact / observation we care about

Why worry about morphology in MT

- Much of statistical NLP:
  estimate and use $p_\theta(y|x)$, $y \in Y$, $x \in X$
- X = representation for language, including words
- Y = some event / fact / observation we care about
- Sparseness: Estimating p_θ is hard because language (event space X) is very sparse.

Why worry about morphology in MT

- Much of statistical NLP:
  estimate and use $p_\theta(y|x)$, $y \in Y$, $x \in X$
- X = representation for language, including words
- Y = some event / fact / observation we care about
- Sparseness: Estimating p_θ is hard because language (event space X) is very sparse.
- Morphological analysis can reduce sparseness.

Why worry about morphology in MT

- Much of statistical NLP:
  estimate and use $p_\theta(y|x)$, $y \in Y$, $x \in X$
- X = representation for language, including words
- Y = some event / fact / observation we care about
- Sparseness: Estimating p_θ is hard because language (event space X) is very sparse.
- Morphological analysis can reduce sparseness.
- Morphological analysis improves estimates of p_θ.

Why worry about morphology in MT

- Much of statistical NLP:
  estimate and use $p_\theta(y|x)$, $y \in Y$, $x \in X$
- X = representation for language, including words
- Y = some event / fact / observation we care about
- Sparseness: Estimating p_θ is hard because language (event space X) is very sparse.
- Morphological analysis can reduce sparseness.
- Morphological analysis improves estimates of p_θ.
- English is morphologically poor,
  so simple heuristics are often sufficient.

Why worry about morphology in MT

- Much of statistical NLP:
  estimate and use $p_\theta(y|x)$, $y \in Y$, $x \in X$
- X = representation for language, including words
- Y = some event / fact / observation we care about
- Sparseness: Estimating p_θ is hard because language (event space X) is very sparse.
- Morphological analysis can reduce sparseness.
- Morphological analysis improves estimates of p_θ.
- English is morphologically poor,
  so simple heuristics are often sufficient.
- Also true for many other languages.

Why worry about morphology in MT

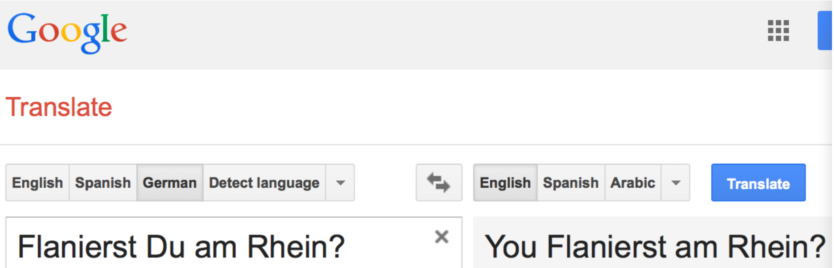
- Much of statistical NLP:
  estimate and use $p_\theta(y|x)$, $y \in Y$, $x \in X$
- X = representation for language, including words
- Y = some event / fact / observation we care about
- Sparseness: Estimating p_θ is hard because language (event space X) is very sparse.
- Morphological analysis can reduce sparseness.
- Morphological analysis improves estimates of p_θ.
- English is morphologically poor,
  so simple heuristics are often sufficient.
- Also true for many other languages.
- So this part of the talk only applies to pairs of languages of which at least one is morphologically rich.

Why worry about morphology in MT

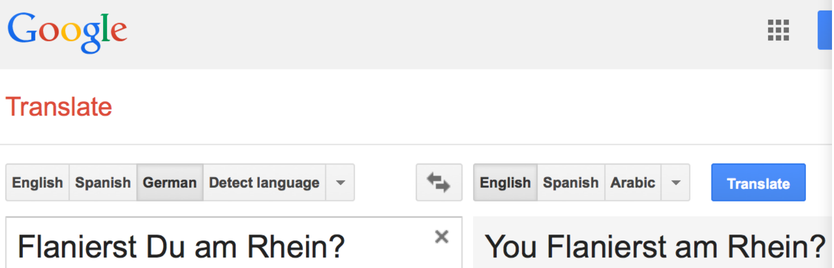
- For symbolic / rule-based approaches, there is a very similar argument for why you need morphology if you are dealing with a morphologically rich language.

Anecdotal example

Anecdotal example

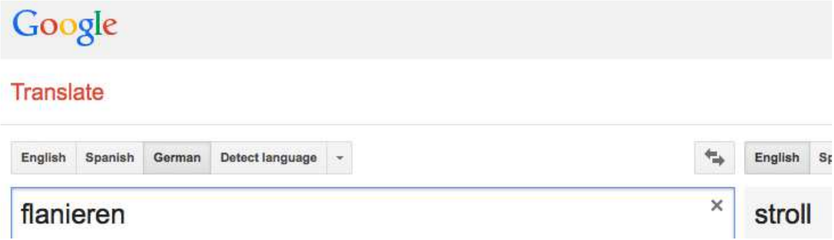


Anecdotal example



Inflected form "flanierst" is not translated.

Anecdotal example



The lemma "flanieren" is correctly translated as "to stroll".

Why worry about morphology now?

Why worry about morphology now?

- Recent progress: new technology for high-accuracy high-performance morphological analysis

Why worry about morphology now?

- Recent progress: new technology for high-accuracy high-performance morphological analysis
- Resources (linguistically annotated corpora) are becoming available for an increasing number of languages.

http://cistern.cis.lmu.de/marmot

cistern.cis.lmu.de/marmot/

CIS projects: SMOR, MarMoT, Ocrocis, SFST, MarLiN, complete list

MarMoT - A fast and accurate morphological tagger

(Source: wikimedia.org)

MarMoT is a generic conditional random field (CRF) framework as well as a state-of-the-art morphological tagger.
On this page you can find links to the source code, binaries, pretrained models, automatically annotated datasets and more.

MarMoT model for German (freely available)

Sei	sein	number=sg|person=3|tense=pres|mo
diese	dieser	case=nom|number=sg|gender=fem
überschritten	überschreiten	_
,	,	_
würden	werden	number=pl|person=3|tense=past|mo
die	der	case=nom|number=pl|gender=*
‘	‘	_
Signale	signal	case=nom|number=pl|gender=neut
nicht	nicht	_
hart	hart	degree=pos
gestellt	stellen	_
”	”	_
.	.	_

MarMoT model for Czech (freely available)

Názor	názor	num=s|gen=m|cas=a
experta	expert	num=s|gen=m|cas=a
Informace	informace	num=p|gen=f|cas=n
zveřejněné	zveřejněný	num=p|gen=f|deg=p|cas=n
v	v	cas=l
Profitu	profit	num=s|gen=m|cas=l
o	o	cas=l
možnostech	možnost	num=p|gen=f|cas=l
využití	využití	num=s|gen=n|cas=n
poradců	poradce	num=p|gen=m|cas=g

MarMoT model for Hungarian (freely available)

A	a	`SubPOS=f`
gazdaság	gazdaság	`SubPOS=c|Num=s|Cas=n|NumP=none|PerP=nor`
ilyen	ilyen	`SubPOS=d|Per=3|Num=s|Cas=n|NumP=none|Pe`
mértékű	mértékű	`SubPOS=f|Deg=p|Num=s|Cas=n|NumP=none|Pe`
fejlődését	fejlődés	`SubPOS=c|Num=s|Cas=a|NumP=s|PerP=3|Num`
több	több	`SubPOS=c|Num=s|Cas=n|Form=l|NumP=none|`
folyamat	folyamat	`SubPOS=c|Num=s|Cas=n|NumP=none|PerP=nor`
gerjeszti	gerjeszti	`SubPOS=f|Deg=p|Num=s|Cas=n|NumP=none|Pe`

MarMoT model for Spanish (freely available)

que	que	`type=r|num=n|gen=c`
se	se	`type=r|num=n|gen=c|per=3`
llamaba	llamar	`type=m|num=s|mood=i|ten=i|per=3`
la	el	`type=a|num=s|gen=f`
voz	voz	`type=c|num=s|gen=f`
de	de	`type=p|form=s`
la	el	`type=a|num=s|gen=f`
conciencia	conciencia	`type=c|num=s|gen=f`

MarMoT model for Latin (freely available)

Cum	cum	INFL=n
autem	autem	INFL=n
perambulasset	perambulo	PERS=3|NUMB=s|TENS=l|MOOD=s|VOIC=a
partes	pars	NUMB=p|GEND=f|CASE=a
illas	ille	NUMB=p|GEND=f|CASE=a
et	et	INFL=n
exhortatus	exhorto	NUMB=s|TENS=r|MOOD=p|VOIC=p|GEND=m
eos	is	PERS=3|NUMB=p|GEND=m|CASE=a

MarMoT model for English (freely available)

The	the	DT
agreements	agreement	NNS
bring	bring	VBP
to	to	IN
a	a	DT
total	total	NN
of	of	IN
nine	nine	CD
the	the	DT
number	number	NN
of	of	IN
planes	plane	NNS
the	the	DT
travel	travel	NN
company	company	NN
has	have	VBZ
sold	sell	VBN

What do I need to train a model for a new language?

What do I need to train a model for a new language?

- A morphologically annotated corpus

What do I need to train a model for a new language?

- A morphologically annotated corpus
  - usually 10,000 to 100,000 tokens if annotation is high quality

What do I need to train a model for a new language?

- A morphologically annotated corpus
  - usually 10,000 to 100,000 tokens if annotation is high quality
  - more in some cases and if the annotation is not high quality

What do I need to train a model for a new language?

- A morphologically annotated corpus
    - usually 10,000 to 100,000 tokens if annotation is high quality
    - more in some cases and if the annotation is not high quality
- Given this resource, training a MarMoT model is efficient and simple.

Results (Müller, Cotterell, Fraser, Schütze, 2015)

				cs		de		en		es		hu		la	
1	PCRF		tag	89.75	76.83	82.81	61.60	**96.45**	**90.68**	97.05	90.07	93.64	84.65	82.37	53.73
2	JCK		lemma	95.95	81.28	96.63	85.84	99.08	94.28	97.69	87.19	96.69	88.66	90.79	58.23
3	JCK		tag+lemma	87.85	67.00	81.60	55.97	96.17	87.32	95.44	80.62	92.15	78.89	79.51	39.07
4	LEMMING-P	+dict	lemma	97.46	89.14	97.70	91.27	**99.21**	**95.59**	98.48	92.98	97.53	92.10	93.07	69.83
5	LEMMING-P	+dict	tag+lemma	88.86	72.51	82.27	59.42	**96.27**	**88.49**	96.12	85.80	92.59	80.77	80.49	44.26
6	LEMMING-P	+mrph	lemma	97.29	88.98	97.51	90.85	NA	NA	98.68	94.32	97.53	92.15	92.54	67.81
7	LEMMING-P	+mrph	tag+lemma	89.23	74.24	82.49	60.42	NA	NA	96.35	87.25	93.11	82.56	80.67	45.21
8	LEMMING-J	+dict	tag	**90.34**$^{+}$	78.47	**83.10**$^{+}$	62.36	96.32	89.70	97.11	90.13	93.64	84.78	82.89	54.69
9	LEMMING-J	+dict	lemma	98.27	92.67	**98.10**$^{+}$	92.79	**99.21**	95.23	98.67	94.07	98.02	94.15	**95.58**$^{+}$	**81.74**$^{+}$
10	LEMMING-J	+dict	tag+lemma	89.69	75.44	82.64	60.49	96.17	87.87	96.23	86.19	92.84	81.89	81.92	49.97
11	LEMMING-J	+mrph	tag	90.20	**79.72***	**83.10**$^{+}$	**63.10***	NA	NA	**97.16**	**90.66**	**93.67**	**85.12**	**83.49***	**58.76***
12	LEMMING-J	+mrph	lemma	**98.42***	**93.46***	**98.10**$^{+}$	**93.02**$^{+}$	NA	NA	**98.78***	**94.86***	**98.08**$^{+}$	**94.26**$^{+}$	95.36	80.94
13	LEMMING-J	+mrph	tag+lemma	**89.90***	**78.34***	**82.84***	**62.10***	NA	NA	**96.41**$^{\times}$	**87.47**$^{\times}$	**93.40***	**84.15***	**82.57**$^{+}$	**54.63**$^{+}$

Results (Müller, Cotterell, Fraser, Schütze, 2015)

l = lemmatization

t/l = taggig and lemmatization

	cs		de		es		hu		la	
	ALL	OOV	ALL	OOV	ALL	OOV	ALL	OOV	ALL	OOV
l	98.42	93.46	98.10	93.02	98.78	94.86	98.08	94.26	95.36	80.94
t/l	89.90	78.34	82.84	62.10	96.41	87.47	93.40	84.15	82.57	54.63

Lemmatization vs. Morphological features

- Lemmatization ready for prime time
- Morphological features: you may need more than 100,000 tokens in some languages

Summary

Summary

- Recent progress in morphological analysis:
  due to new technology and new resources

Summary

- Recent progress in morphological analysis: due to new technology and new resources
- Morphological analysis reduces sparseness.

Summary

- Recent progress in morphological analysis: due to new technology and new resources
- Morphological analysis reduces sparseness.
- $\Rightarrow$ better machine translation

Outline

1. Morphology
2. Deep learning embeddings
3. Morphological lexica vs embeddings
4. For units of which granularities should we use embeddings?
5. Using deep learning (in general) in MT

What are embeddings?

Why use embeddings for MT?

Why use embeddings for MT?

- Discrete space $\Rightarrow$ Continuous space

Why use embeddings for MT?

- Discrete space $\Rightarrow$ Continuous space
- Much of statistical NLP: estimate and use $p_\theta(y|x)$, $x \in X$

Why use embeddings for MT?

- Discrete space $\Rightarrow$ Continuous space
- Much of statistical NLP: estimate and use $p_\theta(y|x)$, $x \in X$
- X = representation for language, including words

Why use embeddings for MT?

- Discrete space $\Rightarrow$ Continuous space
- Much of statistical NLP: estimate and use $p_\theta(y|x)$, $x \in X$
- X = representation for language, including words
- If X is discrete:

Why use embeddings for MT?

- Discrete space $\Rightarrow$ Continuous space
- Much of statistical NLP: estimate and use $p_\theta(y|x)$, $x \in X$
- X = representation for language, including words
- If X is discrete:
  - It is often hard to deal with rare/unseen events x.

Why use embeddings for MT?

- Discrete space $\Rightarrow$ Continuous space
- Much of statistical NLP: estimate and use $p_\theta(y|x)$, $x \in X$
- X = representation for language, including words
- If X is discrete:
  - It is often hard to deal with rare/unseen events x.
  - In language modeling: throw away information, guess (e.g., Kneser-Ney)

Why use embeddings for MT?

- Discrete space $\Rightarrow$ Continuous space
- Much of statistical NLP: estimate and use $p_\theta(y|x)$, $x \in X$
- X = representation for language, including words
- If X is discrete:
  - It is often hard to deal with rare/unseen events x.
  - In language modeling: throw away information, guess (e.g., Kneser-Ney)
- If X is continuous:

Why use embeddings for MT?

- Discrete space $\Rightarrow$ Continuous space
- Much of statistical NLP: estimate and use $p_\theta(y|x)$, $x \in X$
- X = representation for language, including words
- If X is discrete:
  - It is often hard to deal with rare/unseen events x.
  - In language modeling: throw away information, guess (e.g., Kneser-Ney)
- If X is continuous:
  - You may be able to handle rare/unseen events well ...

Why use embeddings for MT?

- Discrete space $\Rightarrow$ Continuous space
- Much of statistical NLP: estimate and use $p_\theta(y|x)$, $x \in X$
- X = representation for language, including words
- If X is discrete:
  - It is often hard to deal with rare/unseen events x.
  - In language modeling: throw away information, guess (e.g., Kneser-Ney)
- If X is continuous:
  - You may be able to handle rare/unseen events well ...
  - ... if the continuous space X is smooth in some sense.

Why use embeddings for MT?

- Discrete space $\Rightarrow$ Continuous space
- Much of statistical NLP: estimate and use $p_\theta(y|x)$, $x \in X$
- X = representation for language, including words
- If X is discrete:
  - It is often hard to deal with rare/unseen events x.
  - In language modeling: throw away information, guess (e.g., Kneser-Ney)
- If X is continuous:
  - You may be able to handle rare/unseen events well ...
  - ... if the continuous space X is smooth in some sense.
  - State of the art in language modeling: continuous space

Best level for embeddings?

Best level for embeddings?

- Standard approach: compute embeddings for word forms

Best level for embeddings?

- Standard approach: compute embeddings for word forms
- However, in most cases, the lemma is nexus at which form-meaning pairing is located.

Best level for embeddings?

- Standard approach: compute embeddings for word forms
- However, in most cases, the lemma is nexus at which form-meaning pairing is located.
- Exceptions:
  air/s, blind/s, custom/s, manner/s, spectacle/s, wood/s

Best level for embeddings?

- Standard approach: compute embeddings for word forms
- However, in most cases, the lemma is nexus at which form-meaning pairing is located.
- Exceptions:
  air/s, blind/s, custom/s, manner/s, spectacle/s, wood/s
- However, this can be seen as just one instance of the general phenomenon of noncompositionality in language.

Best level for embeddings?

- Standard approach: compute embeddings for word forms
- However, in most cases, the lemma is nexus at which form-meaning pairing is located.
- Exceptions:
  air/s, blind/s, custom/s, manner/s, spectacle/s, wood/s
- However, this can be seen as just one instance of the general phenomenon of noncompositionality in language.
- E.g., hot dog, red herring, kick the bucket

Best level for embeddings?

- Standard approach: compute embeddings for word forms
- However, in most cases, the lemma is nexus at which form-meaning pairing is located.
- Exceptions:
  air/s, blind/s, custom/s, manner/s, spectacle/s, wood/s
- However, this can be seen as just one instance of the general phenomenon of noncompositionality in language.
- E.g., hot dog, red herring, kick the bucket
- If we pick a single level for embeddings,
  then the lemma level is a good one.

Best level for embeddings?

Best level for embeddings?

Standard approach now: word forms

Best level for embeddings?

Standard approach now: word forms

Wuerden	die	Signale	nicht	hart	gestellt
$\vec{v}_{\text{wuerden}}$	$\vec{v}_{\text{die}}$	$\vec{v}_{\text{Signale}}$	$\vec{v}_{\text{nicht}}$	$\vec{v}_{\text{hart}}$	$\vec{v}_{\text{gestellt}}$

Best level for embeddings?

Standard approach now: word forms

Wuerden	die	Signale	nicht	hart	gestellt
$\vec{v}_{\text{wuerden}}$	$\vec{v}_{\text{die}}$	$\vec{v}_{\text{Signale}}$	$\vec{v}_{\text{nicht}}$	$\vec{v}_{\text{hart}}$	$\vec{v}_{\text{gestellt}}$

Better approach: lemmata

Best level for embeddings?

Standard approach now: word forms

Wuerden	die	Signale	nicht	hart	gestellt
$\vec{v}_{\text{wuerden}}$	$\vec{v}_{\text{die}}$	$\vec{v}_{\text{Signale}}$	$\vec{v}_{\text{nicht}}$	$\vec{v}_{\text{hart}}$	$\vec{v}_{\text{gestellt}}$

Better approach: lemmata

Wuerden	die	Signale	nicht	hart	gestellt
$\vec{v}_{\text{werden}}$	$\vec{v}_{\text{der}}$	$\vec{v}_{\text{signal}}$	$\vec{v}_{\text{nicht}}$	$\vec{v}_{\text{hart}}$	$\vec{v}_{\text{stellen}}$

Best level for embeddings?

Standard approach now: word forms

Wuerden	die	Signale	nicht	hart	gestellt
$\vec{v}_{\text{wuerden}}$	$\vec{v}_{\text{die}}$	$\vec{v}_{\text{Signale}}$	$\vec{v}_{\text{nicht}}$	$\vec{v}_{\text{hart}}$	$\vec{v}_{\text{gestellt}}$

Better approach: lemmata

Wuerden	die	Signale	nicht	hart	gestellt
$\vec{v}_{\text{werden}}$	$\vec{v}_{\text{der}}$	$\vec{v}_{\text{signal}}$	$\vec{v}_{\text{nicht}}$	$\vec{v}_{\text{hart}}$	$\vec{v}_{\text{stellen}}$

Or perhaps: lemmata + morph vectors

Best level for embeddings?

Standard approach now: word forms

Wuerden	die	Signale	nicht	hart	gestellt
$\vec{v}_{\text{wuerden}}$	$\vec{v}_{\text{die}}$	$\vec{v}_{\text{Signale}}$	$\vec{v}_{\text{nicht}}$	$\vec{v}_{\text{hart}}$	$\vec{v}_{\text{gestellt}}$

Better approach: lemmata

Wuerden	die	Signale	nicht	hart	gestellt
$\vec{v}_{\text{werden}}$	$\vec{v}_{\text{der}}$	$\vec{v}_{\text{signal}}$	$\vec{v}_{\text{nicht}}$	$\vec{v}_{\text{hart}}$	$\vec{v}_{\text{stellen}}$

Or perhaps: lemmata + morph vectors

Wuerden	die	Signale	nicht
$\vec{v}_{\text{werden}}\ \vec{v}_{\mu 1}\vec{v}_{\mu 5}\ \ldots$	$\vec{v}_{\text{der}}\ \vec{v}_{\mu 8}\vec{v}_{\mu 1}\ \ldots$	$\vec{v}_{\text{signal}}\ \vec{v}_{\mu 6}\vec{v}_{\mu 2}\ \ldots$	$\vec{v}_{\text{nicht}}\ \vec{v}_{\mu 3}\vec{v}_{\mu 4}$

Best level for embeddings?

Standard approach now: word forms

Wuerden	die	Signale	nicht	hart	gestellt
$\vec{v}_{\text{wuerden}}$	$\vec{v}_{\text{die}}$	$\vec{v}_{\text{Signale}}$	$\vec{v}_{\text{nicht}}$	$\vec{v}_{\text{hart}}$	$\vec{v}_{\text{gestellt}}$

Better approach: lemmata

Wuerden	die	Signale	nicht	hart	gestellt
$\vec{v}_{\text{werden}}$	$\vec{v}_{\text{der}}$	$\vec{v}_{\text{signal}}$	$\vec{v}_{\text{nicht}}$	$\vec{v}_{\text{hart}}$	$\vec{v}_{\text{stellen}}$

Or perhaps: lemmata + morph vectors

Wuerden	die	Signale	nicht
$\vec{v}_{\text{werden}}$ $\vec{v}_{\mu 1} \vec{v}_{\mu 5} \dots$	$\vec{v}_{\text{der}}$ $\vec{v}_{\mu 8} \vec{v}_{\mu 1} \dots$	$\vec{v}_{\text{signal}}$ $\vec{v}_{\mu 6} \vec{v}_{\mu 2} \dots$	$\vec{v}_{\text{nicht}}$ $\vec{v}_{\mu 3} \vec{v}_{\mu 4}$

Or perhaps: lemmata + morph features

Best level for embeddings?

Standard approach now: word forms

Wuerden	die	Signale	nicht	hart	gestellt
$\vec{v}_{\text{wuerden}}$	$\vec{v}_{\text{die}}$	$\vec{v}_{\text{Signale}}$	$\vec{v}_{\text{nicht}}$	$\vec{v}_{\text{hart}}$	$\vec{v}_{\text{gestellt}}$

Better approach: lemmata

Wuerden	die	Signale	nicht	hart	gestellt
$\vec{v}_{\text{werden}}$	$\vec{v}_{\text{der}}$	$\vec{v}_{\text{signal}}$	$\vec{v}_{\text{nicht}}$	$\vec{v}_{\text{hart}}$	$\vec{v}_{\text{stellen}}$

Or perhaps: lemmata + morph vectors

Wuerden	die	Signale	nicht
$\vec{v}_{\text{werden}}\ \vec{v}_{\mu 1}\vec{v}_{\mu 5}\ \ldots$	$\vec{v}_{\text{der}}\ \vec{v}_{\mu 8}\vec{v}_{\mu 1}\ \ldots$	$\vec{v}_{\text{signal}}\ \vec{v}_{\mu 6}\vec{v}_{\mu 2}\ \ldots$	$\vec{v}_{\text{nicht}}\ \vec{v}_{\mu 3}\vec{v}_{\mu 4}$

Or perhaps: lemmata + morph features

Wuerden	die	Signale	nicht	hart
$\vec{v}_{\text{werden}}$ 010010	$\vec{v}_{\text{der}}$ 100010	$\vec{v}_{\text{signal}}$ 111000	$\vec{v}_{\text{nicht}}$ 001100	$\vec{v}_{\text{hart}}$ 00

Summary

- Use embeddings for lemmata, not for word forms

Outline

1. Morphology
2. Deep learning embeddings
3. Morphological lexica vs embeddings
4. For units of which granularities should we use embeddings?
5. Using deep learning (in general) in MT

This section based on work by Thomas Müller.
"Robust Morphological Tagging with Word Representations" (NAACL 2015)

Task: Morphological tagging

Task: Morphological tagging

- Disambiguate part-of-speech and morphology

Task: Morphological tagging

- Disambiguate part-of-speech and morphology
- Example:

Ein	ART	case=nom|number=sg|gender=neut
Klettergebiet	NN	case=nom|number=sg|gender=neut
macht	VVFIN	number=sg|person=3|tense=pres|mood=ind
Geschichte	NN	case=acc|number=sg|gender=fem

Task: Morphological tagging

- Disambiguate part-of-speech and morphology
- Example:

Ein	ART	case=nom|number=sg|gender=neut
Klettergebiet	NN	case=nom|number=sg|gender=neut
macht	VVFIN	number=sg|person=3|tense=pres|mood=ind
Geschichte	NN	case=acc|number=sg|gender=fem

- Part-of-speech disambiguation: ART, NN, VFIN

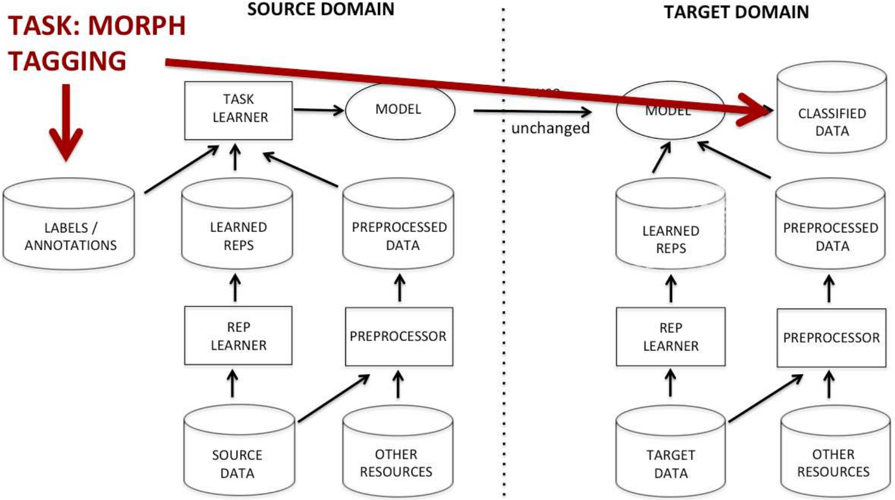
Task: Morphological tagging

- Disambiguate part-of-speech and morphology
- Example:

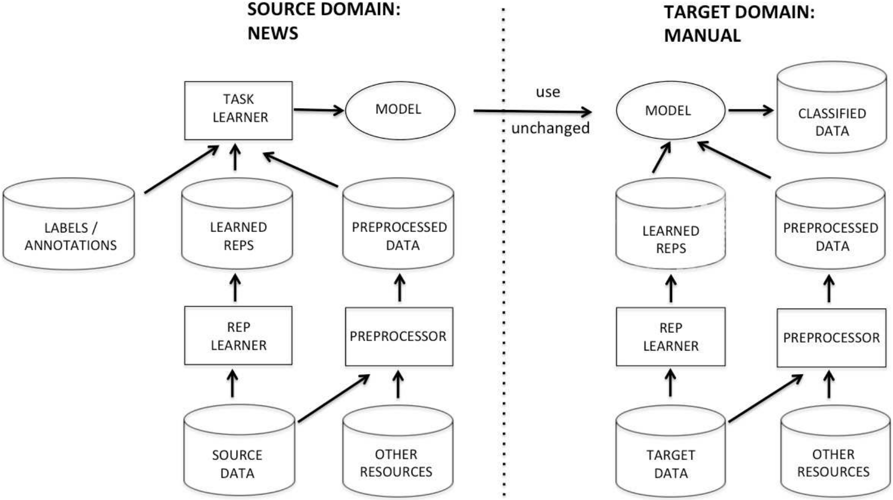
Ein	ART	case=nom|number=sg|gender=neut
Klettergebiet	NN	case=nom|number=sg|gender=neut
macht	VVFIN	number=sg|person=3|tense=pres|mood=ind
Geschichte	NN	case=acc|number=sg|gender=fem

- Part-of-speech disambiguation: ART, NN, VFIN
- Morphological disambiguation: case=nom, number=sg, tense=pres, mood=ind etc

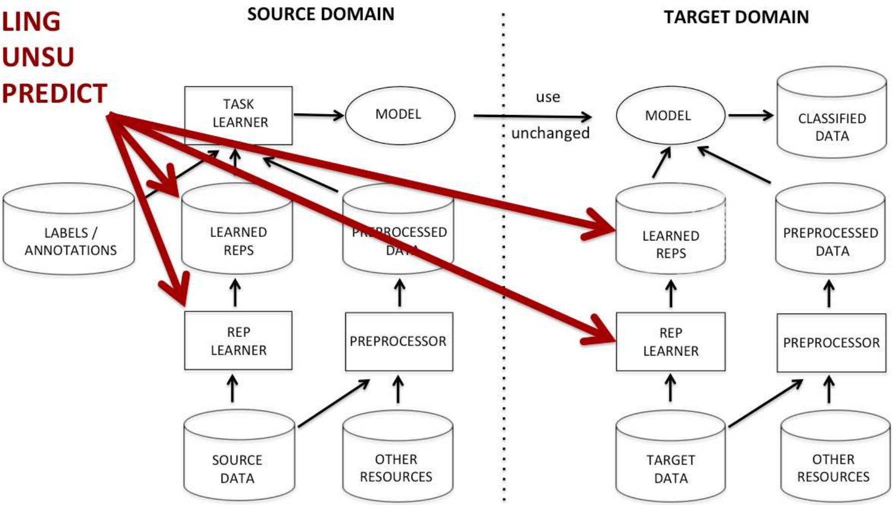
Problem setting: Domain adaptation



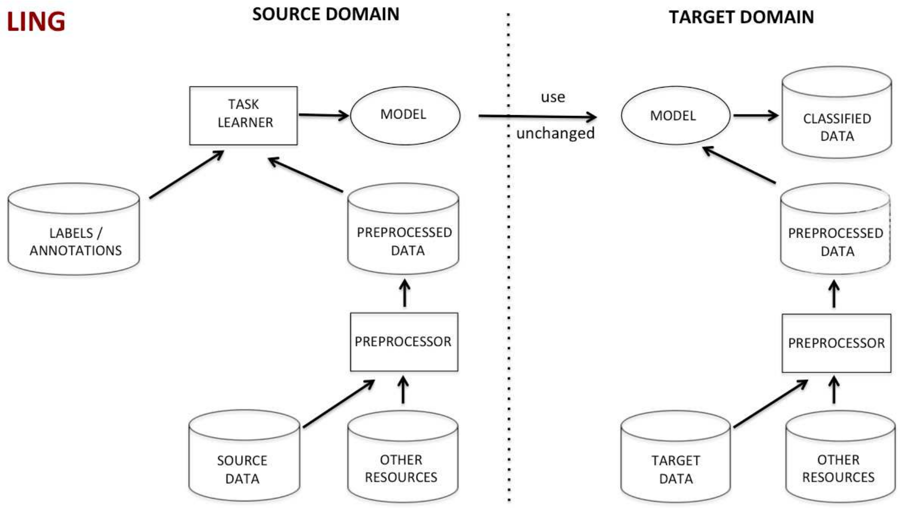
Problem setting: Domain adaptation



Problem setting: Domain adaptation



Problem setting: Domain adaptation



Representation for morphological tagging

Representation for morphological tagging

- Formalize problem as sequence classification
  (using higher-order CRF: MarMoT)

Representation for morphological tagging

- Formalize problem as sequence classification
  (using higher-order CRF: MarMoT)
- Standard features for morphological tagging: suffix, shape, ...

Representation for morphological tagging

- Formalize problem as sequence classification
  (using higher-order CRF: MarMoT)
- Standard features for morphological tagging: suffix, shape, ...
- Additional representation for each token:
  - NONE (word index)
  - UNSU: unsupervised learning: SVD and Brown clusters
  - DEEP: deep learning embeddings
  - LING: finite state morphology (manually created linguistic resource)

Representation for morphological tagging

- Formalize problem as sequence classification (using higher-order CRF: MarMoT)
- Standard features for morphological tagging: suffix, shape, . . .
- Additional representation for each token:
  - NONE (word index)
  - UNSU: unsupervised learning: SVD and Brown clusters
  - DEEP: deep learning embeddings
  - LING: finite state morphology (manually created linguistic resource)
- Which representation works best for morphological tagging: NONE, LING, UNSU or DEEP?

Morphological tagging: Results

	SVMTool	Morfette	MarMoT				
	NONE	NONE	NONE	UNSU1	UNSU2	DEEP	LING
cs	75.28	76.04	78.01	78.44	78.51	78.42	78.88
hu	88.44	89.18	89.77	90.52	90.41	90.88	91.24

Summary

- Embeddings and morphological resources provide complementary information.

Summary

- Embeddings and morphological resources provide complementary information.
- Use both!

Outline

1. Morphology
2. Deep learning embeddings
3. Morphological lexica vs embeddings
4. For units of which granularities should we use embeddings?
5. Using deep learning (in general) in MT

Embeddings for what?

- morphemes
- word forms
- lemmata
- phrases
- sentences
- paragraphs
- documents

Embeddings for what?

Embeddings for what?

- Most common use of embeddings:
  Embeddings for words (= word forms)

Embeddings for what?

- Most common use of embeddings:
  Embeddings for words (= word forms)
- My earlier argument:
  Lemmata are the right level of embedding representation, not word forms.

Embeddings for what?

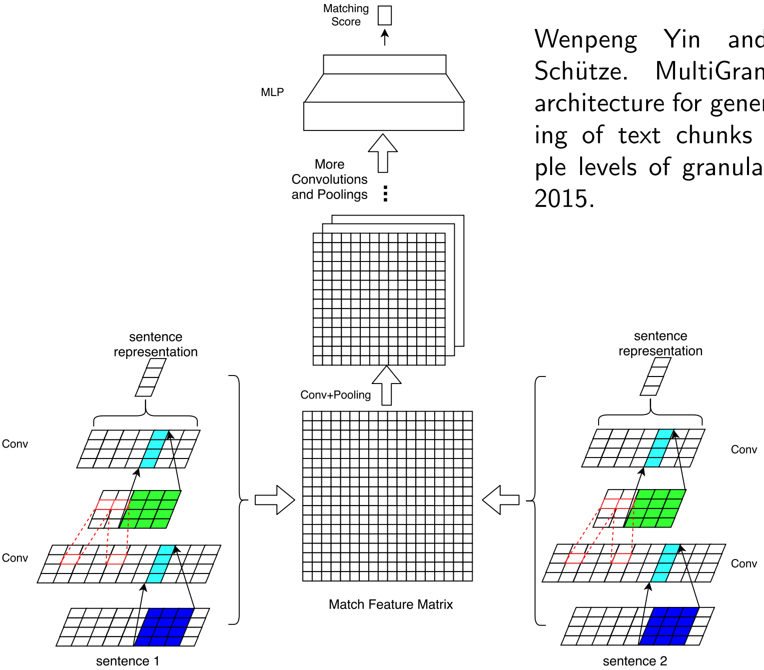
- Most common use of embeddings:
  Embeddings for words (= word forms)
- My earlier argument:
  Lemmata are the right level of embedding representation, not word forms.
- What about embedding representations for larger units: phrases and sentences?

Embeddings for what?

- Most common use of embeddings:
  Embeddings for words (= word forms)
- My earlier argument:
  Lemmata are the right level of embedding representation, not word forms.
- What about embedding representations for larger units: phrases and sentences?
- Recent deep learning work on MT uses
  vector representations for sentences.

Example: paraphrase identification

- Given: two sentences
- Task: Are they paraphrases, yes or no?



Wenpeng Yin and Hinrich Schütze. MultiGranCNN: An architecture for general matching of text chunks on multiple levels of granularity. ACL 2015.

Task-specificity: Experimental results

method	acc	F_1
ARC-I (Hu et al., 2014)	61.4	60.3
ARC-II (Hu et al., 2014)	64.9	63.5
Bi-CNN-MI (Yin and Schütze, 2015)	87.9	87.1
8MT (Madnani et al., 2012)	92.3	92.1
(Bach et al., 2014)	93.4	93.3
MultiGranCNN+8MT (freeze)	94.9	94.7

How to represent sentences: Capacity

- MultiGranCNN determines for each meaning element: is it also present in the other sentence?

How to represent sentences: Capacity

- MultiGranCNN determines for each meaning element:
  is it also present in the other sentence?
- At all levels of granularity:
  single word, short ngram, long ngram, sentence.

How to represent sentences: Capacity

- MultiGranCNN determines for each meaning element: is it also present in the other sentence?
- At all levels of granularity: single word, short ngram, long ngram, sentence.
- Representation of the sentence in MultigranCNN: large set of vectors, each representing a (smaller or larger) part of the sentence.

How to represent sentences: Capacity

- MultiGranCNN determines for each meaning element: is it also present in the other sentence?
- At all levels of granularity: single word, short ngram, long ngram, sentence.
- Representation of the sentence in MultigranCNN: large set of vectors, each representing a (smaller or larger) part of the sentence.
- Alternative:
  - Use a single vector to represent sentence
  - Then compare these two vectors

How to represent sentences: Capacity

- MultiGranCNN determines for each meaning element:
  is it also present in the other sentence?
- At all levels of granularity:
  single word, short ngram, long ngram, sentence.
- Representation of the sentence in MultigranCNN:
  large set of vectors, each representing a (smaller or larger) part of the sentence.
- Alternative:
  - Use a single vector to represent sentence
  - Then compare these two vectors
  - Does it make sense to go through this bottleneck?

How to represent sentences: Capacity

- MultiGranCNN determines for each meaning element: is it also present in the other sentence?
- At all levels of granularity: single word, short ngram, long ngram, sentence.
- Representation of the sentence in MultigranCNN: large set of vectors, each representing a (smaller or larger) part of the sentence.
- Alternative:
  - Use a single vector to represent sentence
  - Then compare these two vectors
  - Does it make sense to go through this bottleneck?
  - Does it make sense to go through this bottleneck for paragraphs?

How to represent sentences: Capacity

- MultiGranCNN determines for each meaning element: is it also present in the other sentence?
- At all levels of granularity: single word, short ngram, long ngram, sentence.
- Representation of the sentence in MultigranCNN: large set of vectors, each representing a (smaller or larger) part of the sentence.
- Alternative:
  - Use a single vector to represent sentence
  - Then compare these two vectors
  - Does it make sense to go through this bottleneck?
  - Does it make sense to go through this bottleneck for paragraphs?
  - Does it make sense to go through this bottleneck for books?

How to represent sentences: Capacity

- MultiGranCNN determines for each meaning element: is it also present in the other sentence?
- At all levels of granularity: single word, short ngram, long ngram, sentence.
- Representation of the sentence in MultigranCNN: large set of vectors, each representing a (smaller or larger) part of the sentence.
- Alternative:
  - Use a single vector to represent sentence
  - Then compare these two vectors
  - Does it make sense to go through this bottleneck?
  - Does it make sense to go through this bottleneck for paragraphs?
  - Does it make sense to go through this bottleneck for books?
- Argument 1 against representing sentences as vectors: Vectors have limited storage capacity.

How to represent sentences: Context

How to represent sentences: Context

- (1) “They continued their advance.”

How to represent sentences: Context

- (1) “They continued their advance.”
- (2) “Houthi forces continued their advance.”
- (3) “Stocks continued their advance”

How to represent sentences: Context

- (1) "They continued their advance."
- (2) "Houthi forces continued their advance."
- (3) "Stocks continued their advance"
- In context, it will be clear that "they" refers either to soldiers or to stocks.

How to represent sentences: Context

- (1) "They continued their advance."
- (2) "Houthi forces continued their advance."
- (3) "Stocks continued their advance"
- In context, it will be clear that "they" refers either to soldiers or to stocks.
- Argument 2 against representing sentences as vectors:
  The same sentence should have different representations in different contexts.

How to represent sentences: Intent

How to represent sentences: Intent

It's impossible to find parking!

It's impossible to find parking!

It's impossible to find parking!

How to represent sentences: Intent

- Why did you not pick up the dry cleaning? –
  It's impossible to find parking!

It's impossible to find parking!

It's impossible to find parking!

How to represent sentences: Intent

- Why did you not pick up the dry cleaning? –
  It's impossible to find parking! (10 minutes ago, it was impossible to find parking at my dry cleaner's.)

It's impossible to find parking!

It's impossible to find parking!

How to represent sentences: Intent

- Why did you not pick up the dry cleaning? –
  It's impossible to find parking! (10 minutes ago, it was impossible to find parking at my dry cleaner's.)
- You're looking for an apartment. Why are you not considering neighborhood X? –
  It's impossible to find parking!

It's impossible to find parking!

How to represent sentences: Intent

- Why did you not pick up the dry cleaning? –
  It's impossible to find parking! (10 minutes ago, it was impossible to find parking at my dry cleaner's.)
- You're looking for an apartment. Why are you not considering neighborhood X? –
  It's impossible to find parking! (It is probably possible to find parking in neighborhood X, but it's difficult, expensive, time-consuming.)

It's impossible to find parking!

How to represent sentences: Intent

- Why did you not pick up the dry cleaning? –
  It's impossible to find parking! (10 minutes ago, it was impossible to find parking at my dry cleaner's.)
- You're looking for an apartment. Why are you not considering neighborhood X? –
  It's impossible to find parking! (It is probably possible to find parking in neighborhood X, but it's difficult, expensive, time-consuming.)
- Why are you late? –
  It's impossible to find parking!

How to represent sentences: Intent

- Why did you not pick up the dry cleaning? –
  It's impossible to find parking! (10 minutes ago, it was impossible to find parking at my dry cleaner's.)
- You're looking for an apartment. Why are you not considering neighborhood X? –
  It's impossible to find parking! (It is probably possible to find parking in neighborhood X, but it's difficult, expensive, time-consuming.)
- Why are you late? –
  It's impossible to find parking! (It actually was not impossible to find parking, it just took a while.)

How to represent sentences: Intent

- Why did you not pick up the dry cleaning? –
  It's impossible to find parking! (10 minutes ago, it was impossible to find parking at my dry cleaner's.)
- You're looking for an apartment. Why are you not considering neighborhood X? –
  It's impossible to find parking! (It is probably possible to find parking in neighborhood X, but it's difficult, expensive, time-consuming.)
- Why are you late? –
  It's impossible to find parking! (It actually was not impossible to find parking, it just took a while.)

Argument 3 against representing sentences as vectors:
Intended meaning depends on communicative task / goal.

Representing a sentence as a vector: Problems

- Capacity
- Representation is context-dependent.
- Representation is task/goal/intent-dependent.

Embeddings for what?

- morphemes
- word forms
- lemmata
- phrases
- sentences
- paragraphs
- documents

Embeddings for what?

- morphemes
- word forms
- lemmata
- phrases
- sentences
- paragraphs
- documents

Embeddings for what?

- morphemes
- word forms
- lemmata
- phrases
- ~~sentences~~
- ~~paragraphs~~
- ~~documents~~

Outline

1. Morphology
2. Deep learning embeddings
3. Morphological lexica vs embeddings
4. For units of which granularities should we use embeddings?
5. Using deep learning (in general) in MT

Deep learning

- Will deep-learning-based MT replace current approaches to MT?
- Yann LeCun, Yoshua Bengio, Geoffrey Hinton: Deep learning. 2015. Nature, 521, 436–444.

Comments on “Deep learning” by LeCun, Bengio & Hinton

Comments on “Deep learning” by LeCun, Bengio & Hinton

On embeddings

Comments on "Deep learning" by LeCun, Bengio & Hinton

On embeddings

N-grams treat each word as an atomic unit, so they cannot generalize across semantically related sequences of words, whereas neural language models can because they associate each word with a vector of real valued features . . .

(thumbs up)

Comments on “Deep learning” by LeCun, Bengio & Hinton

On the “deepness” of deep learning

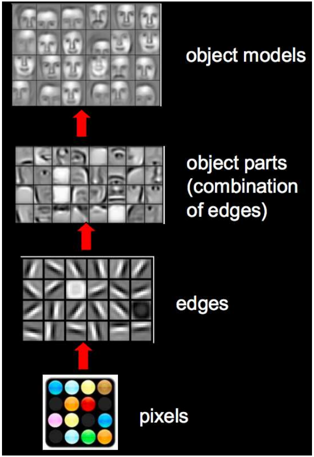
Comments on "Deep learning" by LeCun, Bengio & Hinton

On the "deepness" of deep learning

Deep-learning methods are representation-learning methods with multiple levels of representation, obtained by composing simple but non-linear modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level. With the composition of enough such transformations, very complex functions can be learned.

(thumbs up)

Deep network, increasingly abstract representations



Honglak Lee, Roger Grosse, Rajesh Ranganath, Andrew Y. Ng. 2009. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. ICML 2009.

Comments on "Deep learning" by LeCun, Bengio & Hinton

On convolutional neural networks (CNNs / ConvNets)

Comments on "Deep learning" by LeCun, Bengio & Hinton

On convolutional neural networks (CNNs / ConvNets)

... four key ideas ... local connections, shared weights, pooling and the use of many layers. ... ConvNets have been applied with great success ...

(thumbs up)

Comments on “Deep learning” by LeCun, Bengio & Hinton

Domain expertise no longer needed?

Comments on "Deep learning" by LeCun, Bengio & Hinton

Domain expertise no longer needed?

. . . constructing a pattern-recognition or machine-learning system required careful engineering and considerable domain expertise to design a feature extractor that transformed the raw data (such as the pixel values of an image) into a suitable internal representation . . . deep learning . . . requires very little engineering by hand . . .

(shock)

Comments on "Deep learning" by LeCun, Bengio & Hinton

On unsupervised learning

Comments on “Deep learning” by LeCun, Bengio & Hinton

On unsupervised learning

Although we have not focused on it in this Review, we expect unsupervised learning to become far more important in the longer term. Human and animal learning is largely unsupervised: we discover the structure of the world by observing it . . .

(shock)

Comments on “Deep learning” by LeCun, Bengio & Hinton

On recurrent neural networks (RNNs)

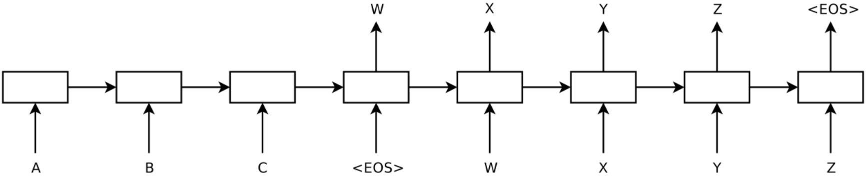
Comments on "Deep learning" by LeCun, Bengio & Hinton

On recurrent neural networks (RNNs)

For tasks that involve sequential inputs, such as speech and language, it is often better to use RNNs . . . RNNs process an input sequence one element at a time, maintaining in their hidden units a 'state vector' that implicitly contains information about the history of all the past elements of the sequence.

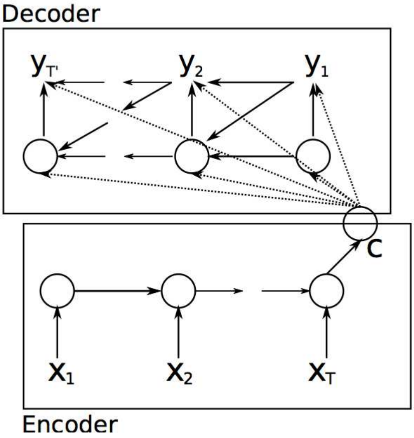
(skepticism – my earlier argument against sentence representation)

Representing sentences as vectors (1)



Sutskever, Vinyals, Le (2015)

Representing sentences as vectors (2)



Cho, Merrienboer, Gulcehre, Bahdanau, Bougares, Schwenk, Bengio (2015)

Comments on "Deep learning" by LeCun, Bengio & Hinton

On symbolic representation / symbolic computation

Comments on “Deep learning” by LeCun, Bengio & Hinton

On symbolic representation / symbolic computation

This rather naive way of performing machine translation has quickly become competitive with the state-of-the-art, and this raises serious doubts about whether understanding a sentence requires anything like the internal symbolic expressions that are manipulated by using inference rules.

(skepticism)

No symbolic representations?

No symbolic representations?

- Compare two types of inference

No symbolic representations?

- Compare two types of inference
  - Memory inference: Inference for frequently observed events, based on retrieval from memory

No symbolic representations?

- Compare two types of inference
  - Memory inference: Inference for frequently observed events, based on retrieval from memory
  - Statistical inference: Inference for never observed events, based on true generalization

No symbolic representations?

- Compare two types of inference
  - Memory inference: Inference for frequently observed events, based on retrieval from memory
  - Statistical inference: Inference for never observed events, based on true generalization
- Example for memory inference

No symbolic representations?

- Compare two types of inference
  - Memory inference: Inference for frequently observed events, based on retrieval from memory
  - Statistical inference: Inference for never observed events, based on true generalization
- Example for memory inference
  - "In Rome, I got a job teaching English as a foreign . . ."

No symbolic representations?

- Compare two types of inference
  - Memory inference: Inference for frequently observed events, based on retrieval from memory
  - Statistical inference: Inference for never observed events, based on true generalization
- Example for memory inference
  - "In Rome, I got a job teaching English as a foreign . . . "
  - What is the next word?

No symbolic representations?

- Compare two types of inference
  - Memory inference: Inference for frequently observed events, based on retrieval from memory
  - Statistical inference: Inference for never observed events, based on true generalization
- Example for memory inference
  - "In Rome, I got a job teaching English as a foreign ..."
  - What is the next word?
  - "language"

No symbolic representations?

- Compare two types of inference
  - Memory inference: Inference for frequently observed events, based on retrieval from memory
  - Statistical inference: Inference for never observed events, based on true generalization
- Example for memory inference
  - "In Rome, I got a job teaching English as a foreign . . ."
  - What is the next word?
  - "language"
- Example for statistical inference

No symbolic representations?

- Compare two types of inference
  - Memory inference: Inference for frequently observed events, based on retrieval from memory
  - Statistical inference: Inference for never observed events, based on true generalization
- Example for memory inference
  - "In Rome, I got a job teaching English as a foreign ..."
  - What is the next word?
  - "language"
- Example for statistical inference
  - "My favorite spice for lighting up shrimp is ..."

No symbolic representations?

- Compare two types of inference
  - Memory inference: Inference for frequently observed events, based on retrieval from memory
  - Statistical inference: Inference for never observed events, based on true generalization
- Example for memory inference
  - "In Rome, I got a job teaching English as a foreign . . ."
  - What is the next word?
  - "language"
- Example for statistical inference
  - "My favorite spice for lighting up shrimp is . . ."
  - What is the next word?

No symbolic representations?

- Compare two types of inference
  - Memory inference: Inference for frequently observed events, based on retrieval from memory
  - Statistical inference: Inference for never observed events, based on true generalization
- Example for memory inference
  - "In Rome, I got a job teaching English as a foreign . . ."
  - What is the next word?
  - "language"
- Example for statistical inference
  - "My favorite spice for lighting up shrimp is . . ."
  - What is the next word?
  - "mace", "garlic", "chili", "paprika"

No symbolic representations?

- Compare two types of inference
  - Memory inference: Inference for frequently observed events, based on retrieval from memory
  - Statistical inference: Inference for never observed events, based on true generalization
- Example for memory inference
  - "In Rome, I got a job teaching English as a foreign ..."
  - What is the next word?
  - "language"
- Example for statistical inference
  - "My favorite spice for lighting up shrimp is ..."
  - What is the next word?
  - "mace", "garlic", "chili", "paprika"
  - probably not: "bay leaf", "saffron"; "vanilla", "allspice"

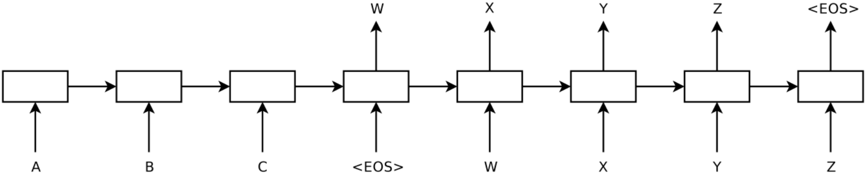
No symbolic representations?

- Compare two types of inference
  - Memory inference: Inference for frequently observed events, based on retrieval from memory
  - Statistical inference: Inference for never observed events, based on true generalization
- Example for memory inference
  - "In Rome, I got a job teaching English as a foreign ..."
  - What is the next word?
  - "language"
- Example for statistical inference
  - "My favorite spice for lighting up shrimp is ..."
  - What is the next word?
  - "mace", "garlic", "chili", "paprika"
  - probably not: "bay leaf", "saffron"; "vanilla", "allspice"
- Is statistical inference the right tool for "memories"?

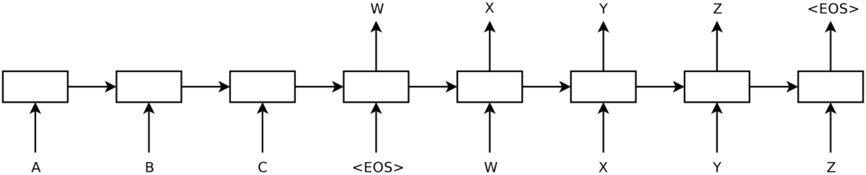
No symbolic representations?

- Compare two types of inference
  - Memory inference: Inference for frequently observed events, based on retrieval from memory
  - Statistical inference: Inference for never observed events, based on true generalization
- Example for memory inference
  - "In Rome, I got a job teaching English as a foreign ..."
  - What is the next word?
  - "language"
- Example for statistical inference
  - "My favorite spice for lighting up shrimp is ..."
  - What is the next word?
  - "mace", "garlic", "chili", "paprika"
  - probably not: "bay leaf", "saffron"; "vanilla", "allspice"
- Is statistical inference the right tool for "memories"?
- Kneser-Ney's success partly due to memorization

Only continuous representations, no symbolic representations?



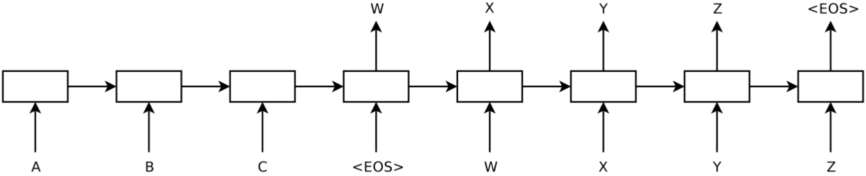
Only continuous representations, no symbolic representations?



nettement
supérieur

Only continuous representations, no symbolic representations?

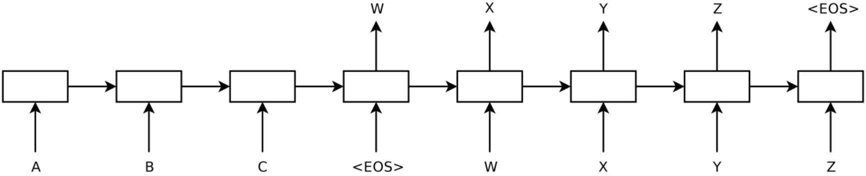
much
higher



nettement
supérieur

Only continuous representations, no symbolic representations?

much
higher

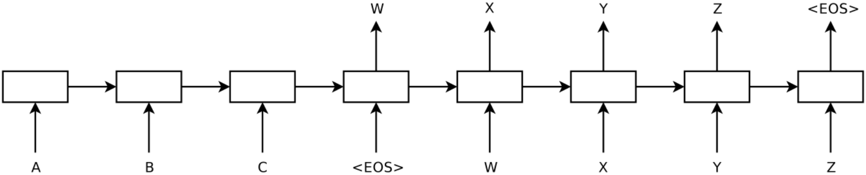


nettement
supérieur

Il lui est nettement supérieur techniquement.

Only continuous representations, no symbolic representations?

much
higher / better

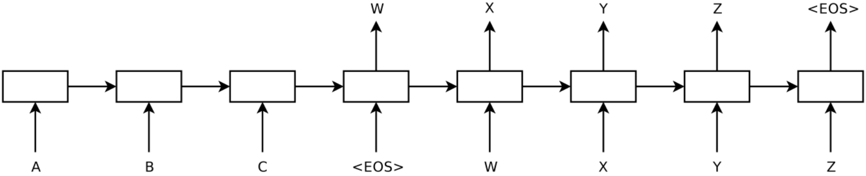


nettement
supérieur

Il lui est nettement supérieur techniquement.

Only continuous representations, no symbolic representations?

much
higher / better

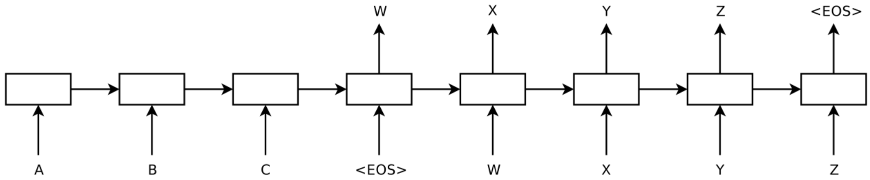


nettement
supérieur

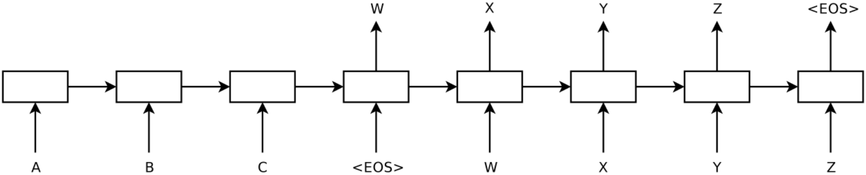
The advantage of a continuous space model

A continuous space model can better learn when to use “higher” vs. “better”.

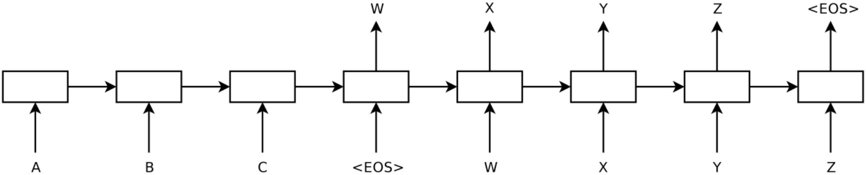
Only continuous representations, no symbolic representations?



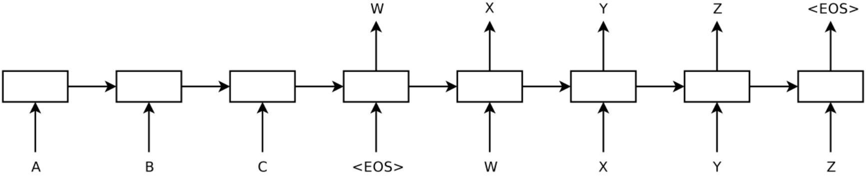
Only continuous representations, no symbolic representations?



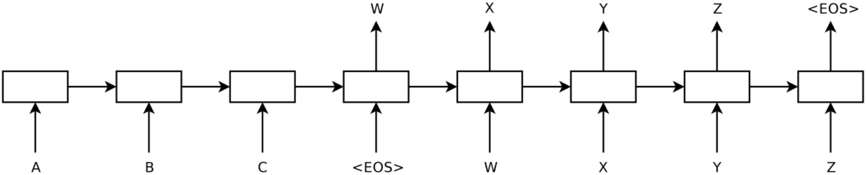
Only continuous representations, no symbolic representations?



Only continuous representations, no symbolic representations?

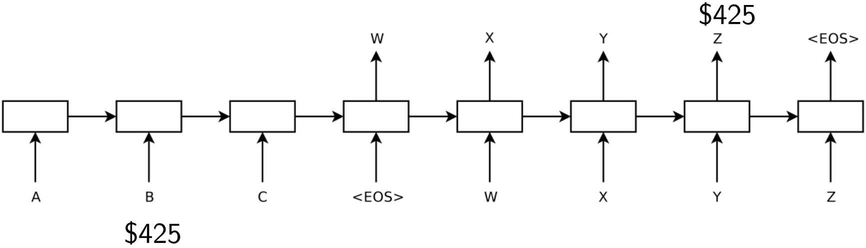


Only continuous representations, no symbolic representations?

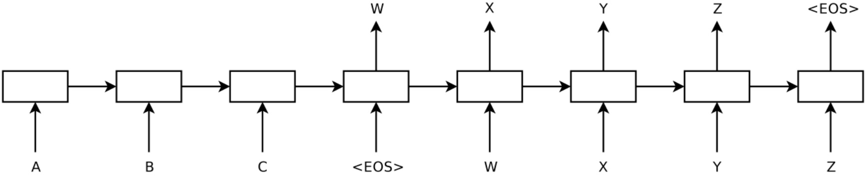


$425

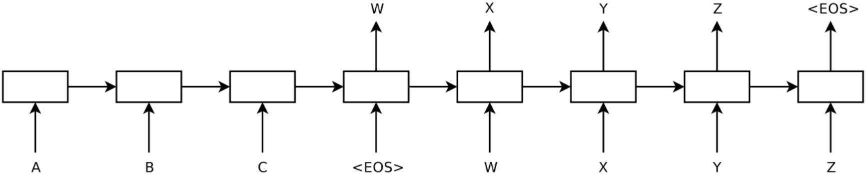
Only continuous representations, no symbolic representations?



Only continuous representations, no symbolic representations?



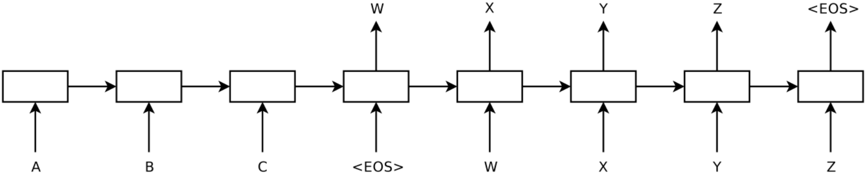
Only continuous representations, no symbolic representations?



5 janvier 1970

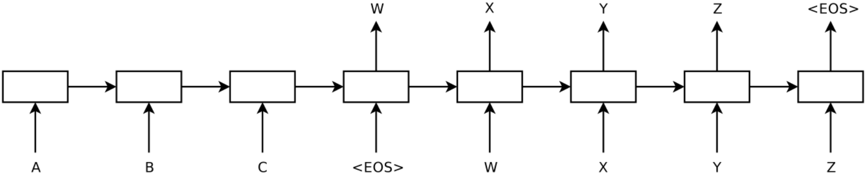
Only continuous representations, no symbolic representations?

Jan 1, 1970



5 janvier 1970

Only continuous representations, no symbolic representations?



Disadvantage of a continuous space model for entities

In translation, it is not a good idea to smooth an entity like Putin, an amount like $425, a date like January 5, 1970.

Summary

Summary

- Use lemmata for MT

Summary

- Use lemmata for MT
- Use embeddings for MT

Summary

- Use lemmata for MT
- Use embeddings for MT
- Use linguistic morphological resources for MT

Summary

- Use lemmata for MT
- Use embeddings for MT
- Use linguistic morphological resources for MT
- Don't represent sentences as vectors for MT

Summary

- Use lemmata for MT
- Use embeddings for MT
- Use linguistic morphological resources for MT
- Don't represent sentences as vectors for MT
- Deep learning will not replace other MT work . . .

Summary

- Use lemmata for MT
- Use embeddings for MT
- Use linguistic morphological resources for MT
- Don't represent sentences as vectors for MT
- Deep learning will not replace other MT work ...
- ... but will be a powerful component of MT systems.