Unsupervised Approaches for Question Answering

Do we really need labeled data?

Cesare Campagnano
campagnano@di.uniroma1.it

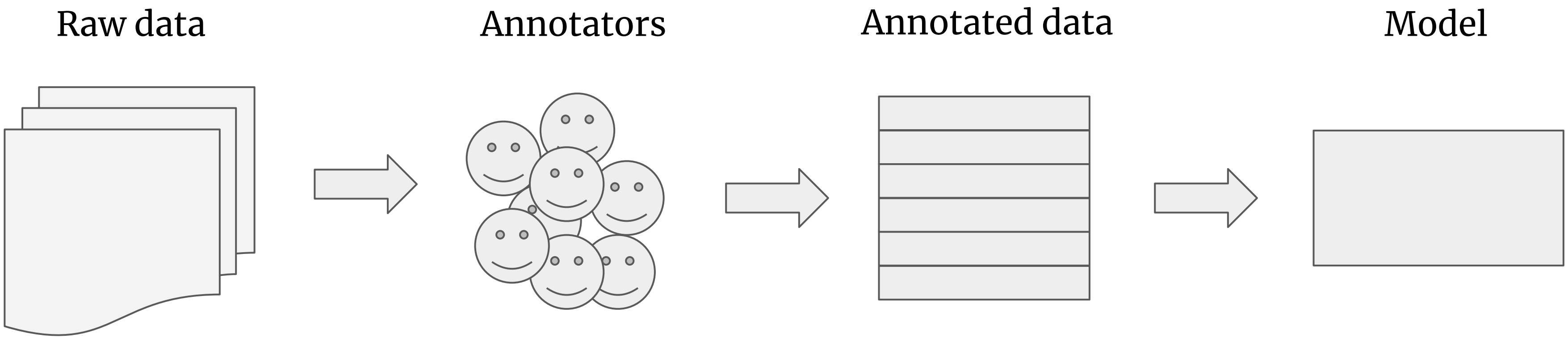
Sapienza Università di Roma

Sapienza NLP

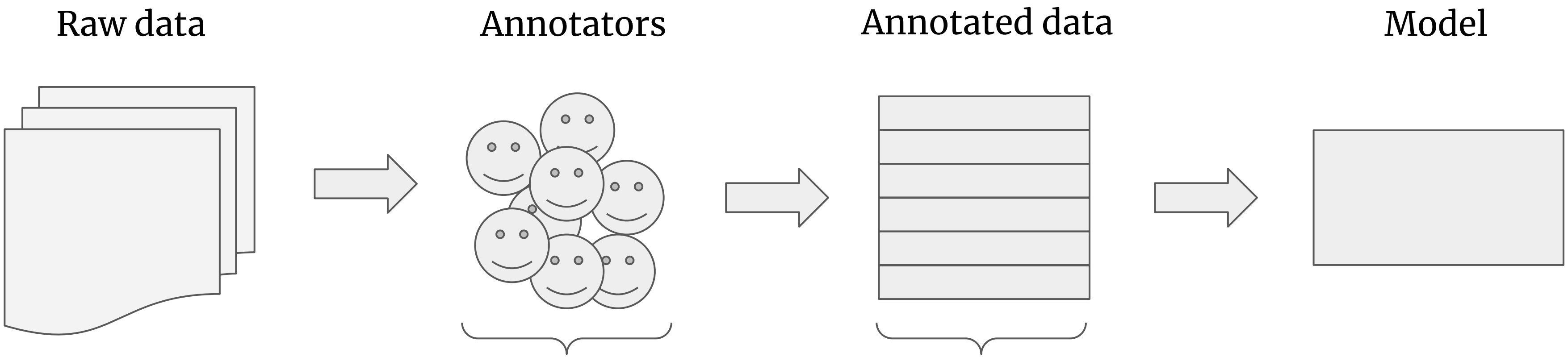
Table of Contents

Introduction

Manually tagging data



Manually tagging data

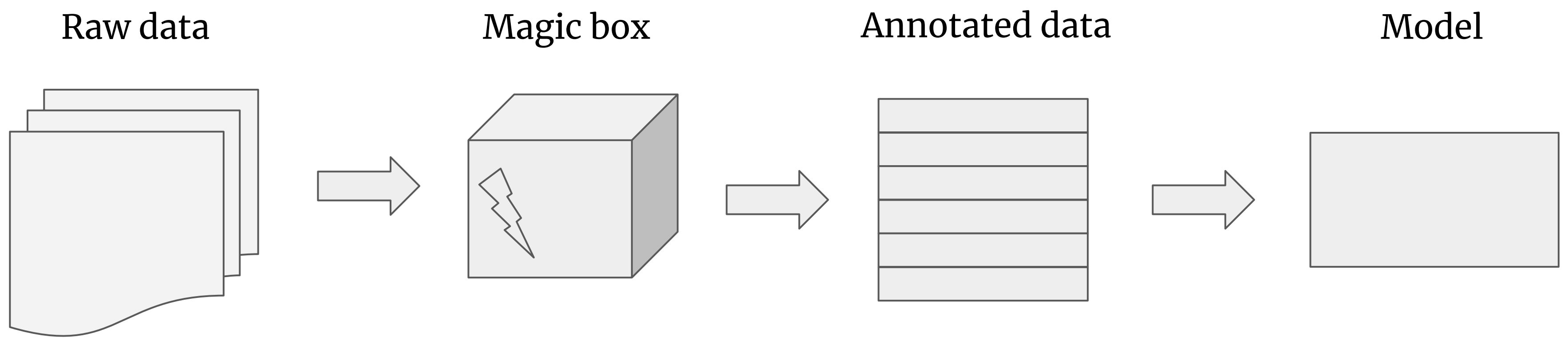


- Expensive
- Time-consuming

- Difficult to scale
- Cannot change domain

SAPIENZA NLP

Unsupervised training data generation



What is *Extractive Question Answering* (EQA)

Question q: `Where did Queen Victoria hold court functions during this time?`

Context c: `Eventually, public opinion forced the Queen to return to London, though even then she preferred to live elsewhere whenever possible. Court functions were still held at Windsor Castle, presided over by the sombre Queen habitually dressed in mourning black, while Buckingham Palace remained shuttered for most of the year.`

What is *Extractive Question Answering* (EQA)

Question q: `Where did Queen Victoria hold court functions during this time?`

Context c: `Eventually, public opinion forced the Queen to return to London, though even then she preferred to live elsewhere whenever possible. Court functions were still held at` **`Windsor Castle`**`, presided over by the sombre Queen habitually dressed in mourning black, while Buckingham Palace remained shuttered for most of the year.`

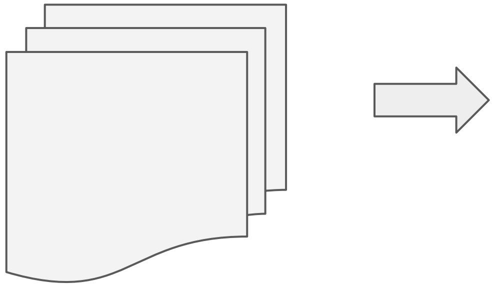
Answer a: `Windsor Castle`

Unsupervised EQA Data Generation

Unsupervised EQA data generation

Unsupervised Question Answering by Cloze Translation (Lewis et al., ACL 2019)

Raw data



$$(c_1, q_1, a_1)$$

$$\vdots$$

$$(c_n, q_n, a_n)$$

Unsupervised EQA data generation

$$P(c, q, a) = \qquad\qquad P(c)$$

Unsupervised EQA data generation

$$P(c, q, a) = \qquad P(a|c) \; P(c)$$

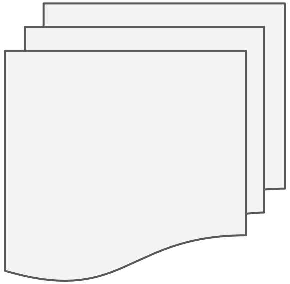
Unsupervised EQA data generation

$$P(c, q, a) = P(q|c, a) \; P(a|c) \; P(c)$$

Unsupervised EQA data generation

$$P(c, q, a) = P(q|c, a)\ P(a|c)\ \boldsymbol{P(c)}$$

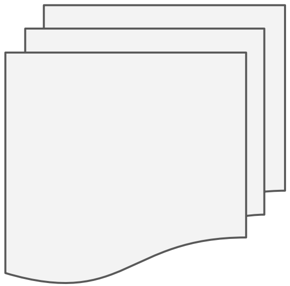
Raw data



Unsupervised EQA data generation

$$P(c, q, a) = P(q|c, a)\ P(a|c)\ \boldsymbol{P(c)}$$

Raw data



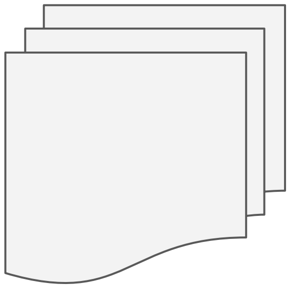
$P(c)$

```
[...] whenever possible.
Court functions were still
held at Windsor Castle,
presided over by the sombre
Queen habitually [...]
```

Unsupervised EQA data generation

$$P(c, q, a) = P(q|c, a)\ \boldsymbol{P(a|c)}\ P(c)$$

Raw data



$P(c)$

```
[...] whenever possible.
Court functions were still
held at Windsor Castle,
presided over by the sombre
Queen habitually [...]
```

NER

$P(a|c)$

`Windsor Castle`

Unsupervised EQA data generation

$$P(c, q, a) = \boldsymbol{P(q|c, a)}\ P(a|c)\ P(c)$$

context c, answer a

```
[...] whenever possible.
Court functions were still
held at Windsor Castle,
presided over by the sombre
Queen habitually [...]
```

Unsupervised EQA data generation

$$P(c, q, a) = \boldsymbol{P(q|c, a)}\ \ P(a|c)\ \ P(c)$$

context c, answer a

```
[...] whenever possible.
Court functions were still
held at Windsor Castle,
presided over by the sombre
Queen habitually [...]
```

cloze question q'

```
[...] whenever possible.
Court functions were still
held at ______________,
presided over by the sombre
Queen habitually [...]
```

Cloze Translation

$q' \Rightarrow q$

- Naïve baseline (identity cloze)

```
Court functions were still held at ______________,
presided over by the sombre Queen [...]
```

[1] Heilman and Smith, 2010
[2] Lample et al., 2018

Cloze Translation

$q' \Rightarrow q$

- Naïve baseline (identity cloze)
- Hard baseline (Noisy cloze)

```
Court functions were still held at ______________,
presided over by the sombre Queen [...]

Where over Court sombre were Queen functions held
at BLANK presided still by the ?
```

[1] Heilman and Smith, 2010
[2] Lample et al., 2018

Cloze Translation

$$q' \Rightarrow q$$

- Naïve baseline (identity cloze) — `Court functions were still held at _____________, presided over by the sombre Queen [...]`
- Hard baseline (Noisy cloze) — `Where over Court sombre were Queen functions held at BLANK presided still by the ?`
- Rule based (Statement-to-question [1]) — `Where Court functions still were held at ?`

[1] Heilman and Smith, 2010
[2] Lample et al., 2018

Cloze Translation

$q' \Rightarrow q$

- Naïve baseline (identity cloze)

  `Court functions were still held at ______________, presided over by the sombre Queen [...]`

- Hard baseline (Noisy cloze)

  `Where over Court sombre were Queen functions held at BLANK presided still by the ?`

- Rule based (Statement-to-question [1])

  `Where Court functions still were held at ?`

- Unsupervised Neural MT [2]

  `Where did sombre Queen still hold Court functions ?`

[1] Heilman and Smith, 2010
[2] Lample et al., 2018

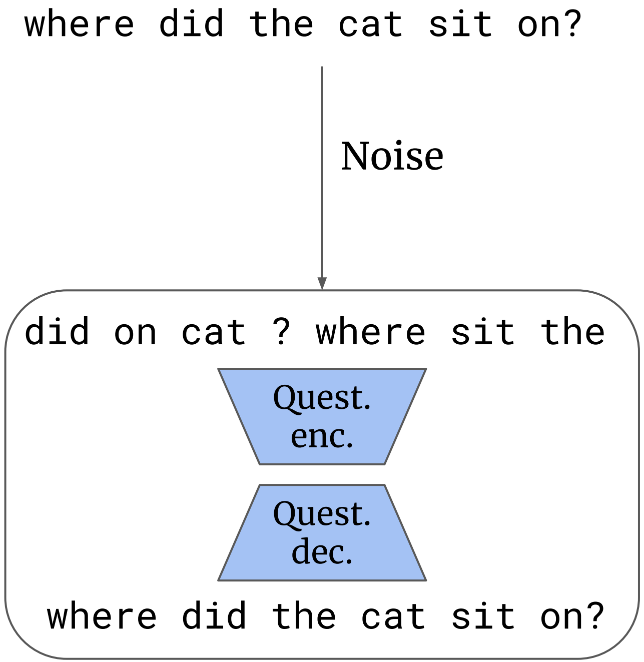
Neural Unsupervised Cloze Translation

Auto-encoder

Back-translation

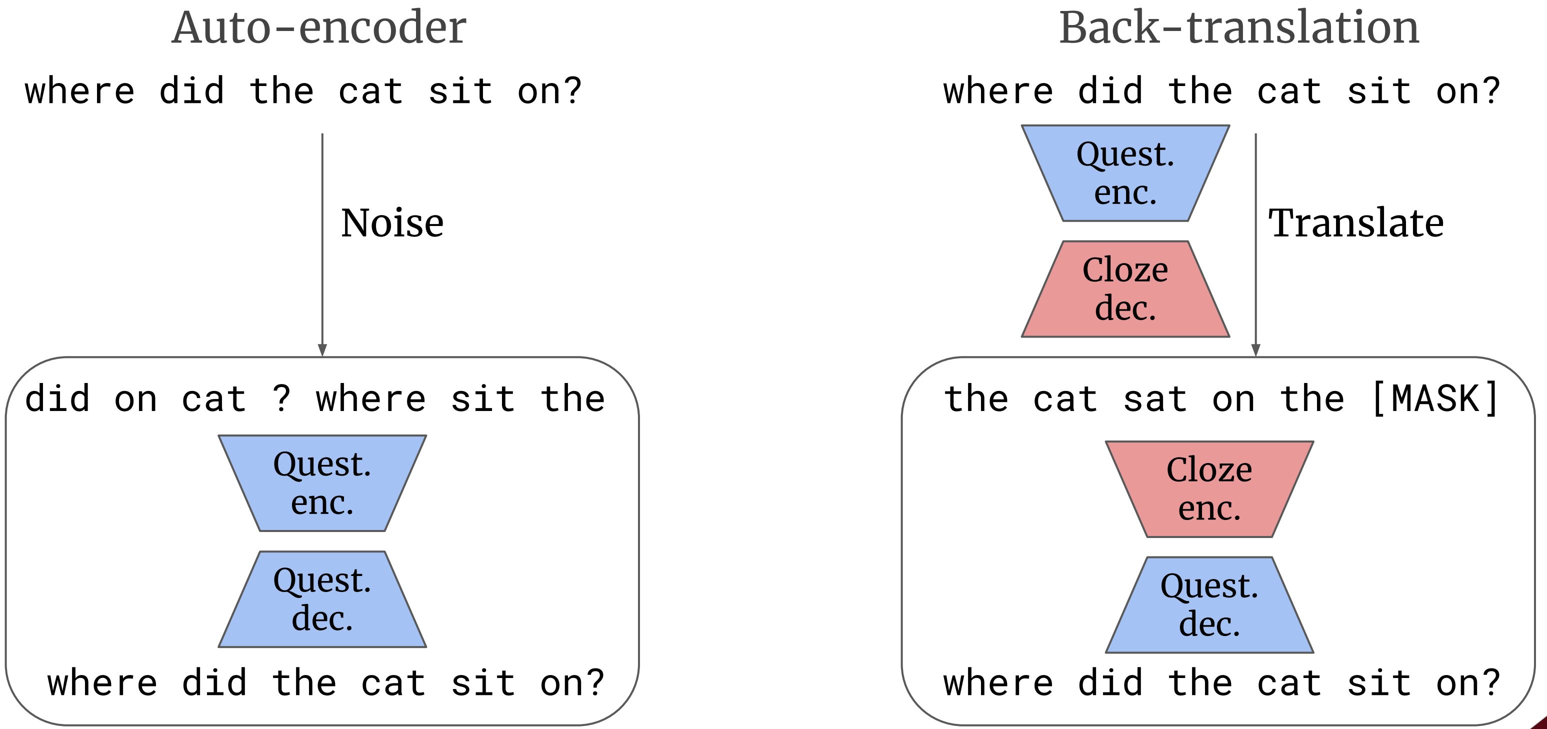
Neural Unsupervised Cloze Translation

Auto-encoder

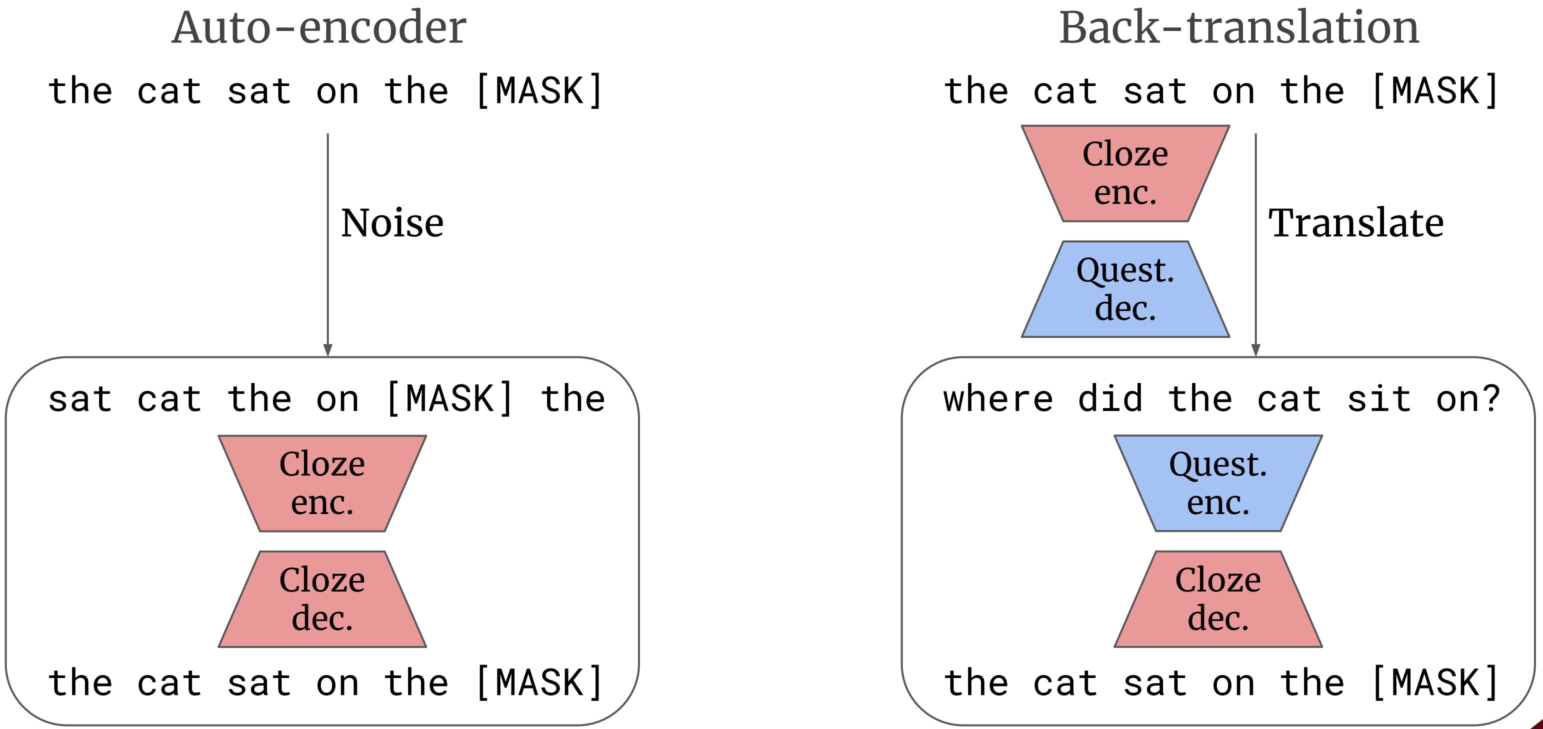


Back-translation

Neural Unsupervised Cloze Translation



Neural Unsupervised Cloze Translation



Unsupervised Cloze Translation Examples

Cloze Question	Answer	Question

Unsupervised Cloze Translation Examples

Cloze Question	Answer	Question
WALA would be sold to the Des Moines-based **ORG** for $86 million	Meredith Corp	Who would buy the WALA Des Moines-based for $86 million?

Unsupervised Cloze Translation Examples

Cloze Question	Answer	Question
WALA would be sold to the Des Moines-based **ORG** for $86 million	Meredith Corp	Who would buy the WALA Des Moines-based for $86 million?
The **NUMERIC** on Orchard Street remained open until 2009	second	How much longer did Orchard Street remain open until 2009?

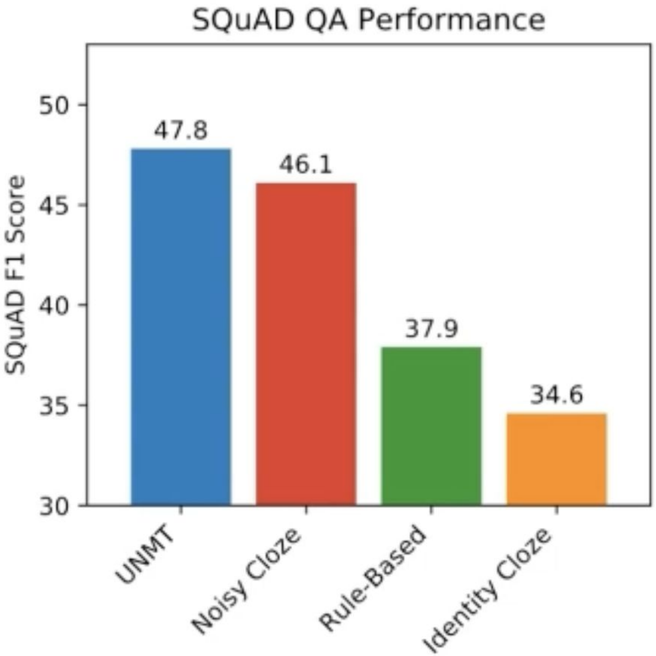
Unsupervised Cloze Translation Examples

Cloze Question	Answer	Question
WALA would be sold to the Des Moines-based **ORG** for $86 million	Meredith Corp	Who would buy the WALA Des Moines-based for $86 million?
The **NUMERIC** on Orchard Street remained open until 2009	second	How much longer did Orchard Street remain open until 2009?
he speaks **LANGUAGE**, English, and German	Spanish	What are we , English , and German?

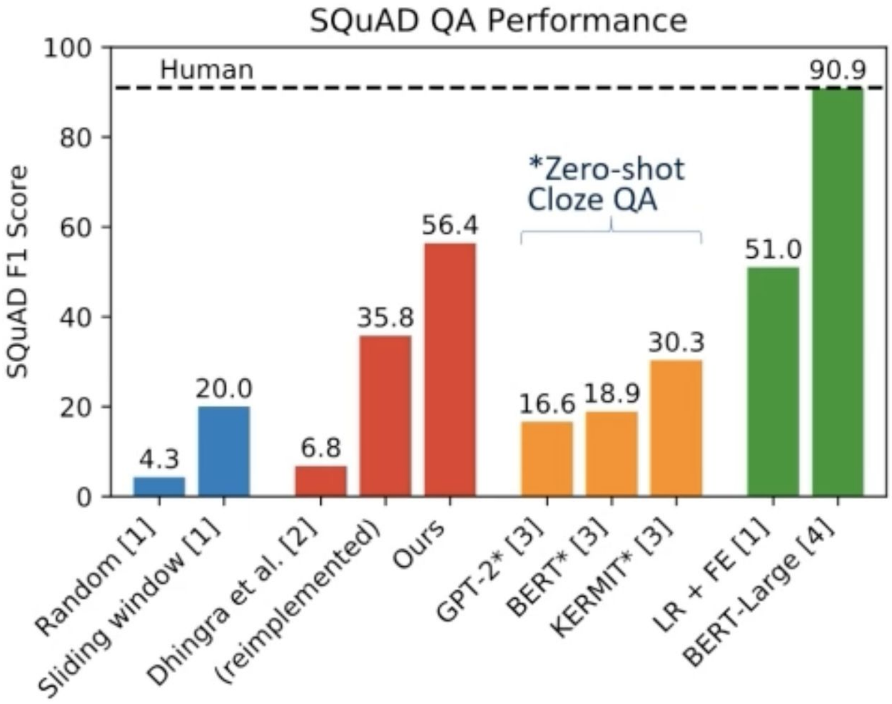
Unsupervised Cloze Translation Examples

Cloze Question	Answer	Question
WALA would be sold to the Des Moines-based **ORG** for $86 million	Meredith Corp	Who would buy the WALA Des Moines-based for $86 million?
The **NUMERIC** on Orchard Street remained open until 2009	second	How much longer did Orchard Street remain open until 2009?
he speaks **LANGUAGE**, English, and German	Spanish	What are we , English , and German?
Form a larger Mid-Ulster District Council in **TEMPORAL**	August	When is a larger Mid-Ulster District Council?

Results


SQuAD QA Performance
SQuAD F1 Score
50
45
40
35
30
47.8
46.1
37.9
34.6
UNMT
Noisy Cloze
Rule-Based
Identity Cloze

Comparison


SQuAD QA Performance
SQuAD F1 Score
100
80
60
40
20
0
Human
90.9
*Zero-shot
Cloze QA
4.3
20.0
6.8
35.8
56.4
16.6
18.9
30.3
51.0
Random [1]
Sliding window [1]
Dhingra et al. [2]
(reimplemented)
Ours
GPT-2* [3]
BERT* [3]
KERMIT* [3]
LR + FE [1]
BERT-Large [4]

Unsupervised Question Decomposition for QA

What is *Extractive Question Answering* (EQA)

Question *q*: `Where did Queen Victoria hold court functions during this time?`

Context *c*: `Eventually, public opinion forced the Queen to return to London, though even then she preferred to live elsewhere whenever possible. Court functions were still held at` **`Windsor Castle`**`, presided over by the sombre Queen habitually dressed in mourning black, while Buckingham Palace remained shuttered for most of the year.`

Answer *a*: `Windsor Castle`

What is *Extractive Question Answering* (EQA)

Question q: `Where did Queen Victoria hold court functions during this time?`

Context c: `Eventually, public opinion forced the Queen to return to London, though even then she preferred to live elsewhere whenever possible. Court functions were still held at` **`Windsor Castle`**`, presided over by the sombre Queen habitually dressed in mourning black, while Buckingham Palace remained shuttered for most of the year.`

Answer a: `Windsor Castle`

"Single-hop" QA

Multi-hop QA

Question q: What profession do H. L. Mencken and Albert Camus have in common?

Context c_4: Henry Louis Mencken (1880 – 1956) was an American journalist, critic and scholar of American English.

Context c_7: Albert Camus (7 November 1913 – 4 January 1960) was a French philosopher, author, and journalist.

Multi-hop QA

Question q: What profession do H. L. Mencken and Albert Camus have in common?

Context c_4: Henry Louis Mencken (1880 – 1956) was an American **journalist**, critic and scholar of American English.

Context c_7: Albert Camus (7 November 1913 – 4 January 1960) was a French philosopher, author, and **journalist**.

Answer a: journalist

Complex problem?

Divide et impera (Divide-and-conquer):

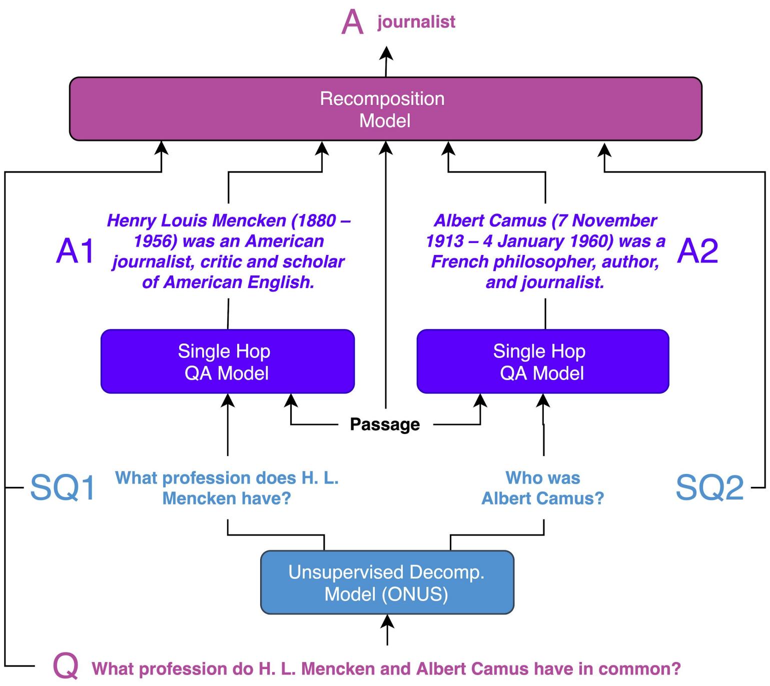
Split hard questions into N simple questions

Unsuperv. Multi-hop Question Decomposition

Unsupervised Question Decomposition

for Question Answering

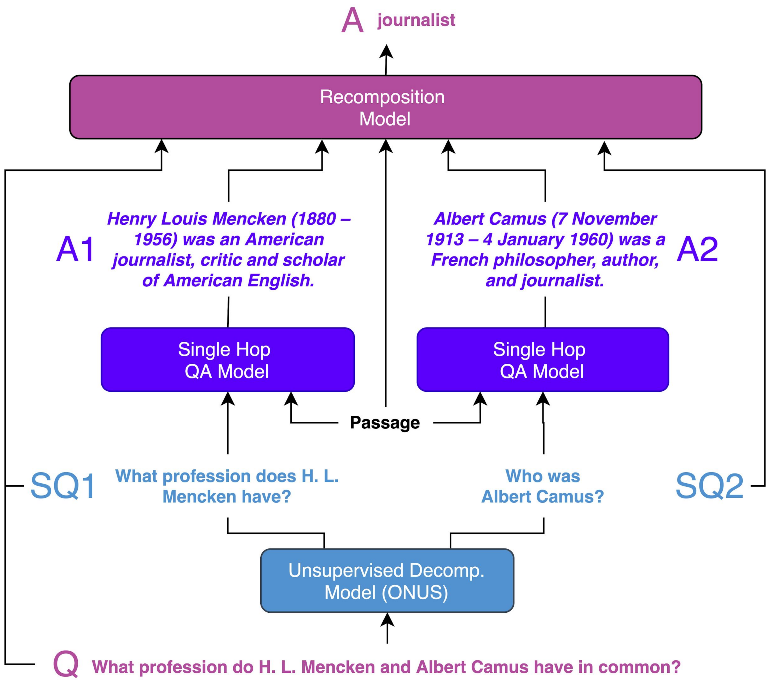
(Perez et al., EMNLP 2020)



Unsuperv. Multi-hop Question Decomposition

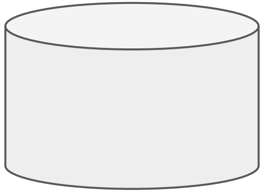
Recomposition model:

$$P(a|c, q, [s_1, a_1], \ldots, [a_N, s_N])$$



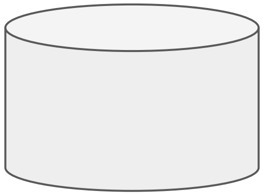
Question corpus creation

Large corpus
of questions



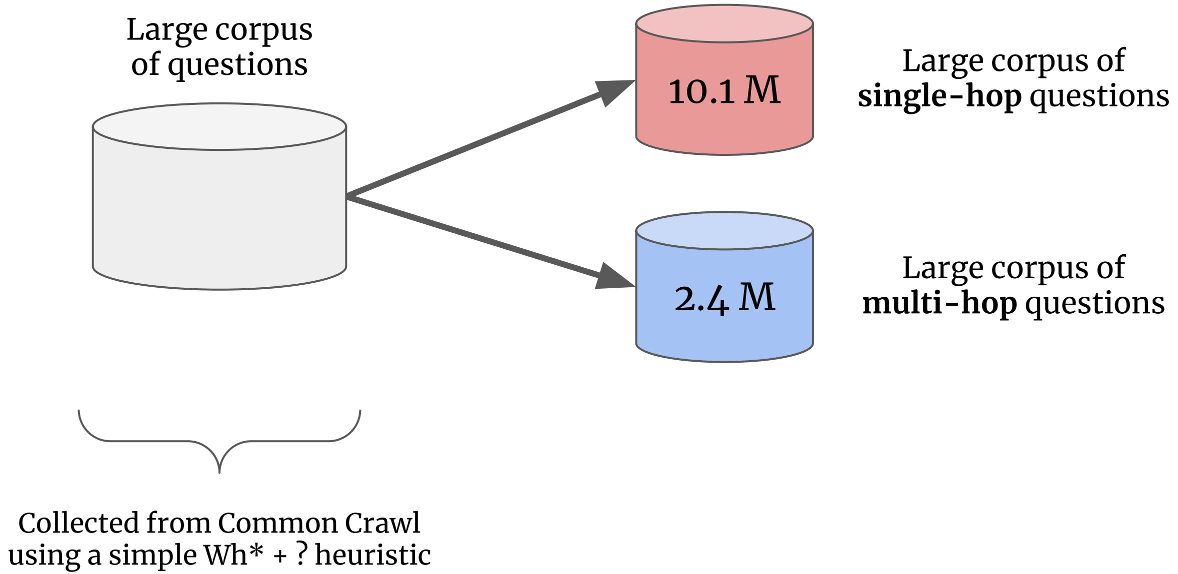
Question corpus creation

Large corpus
of questions



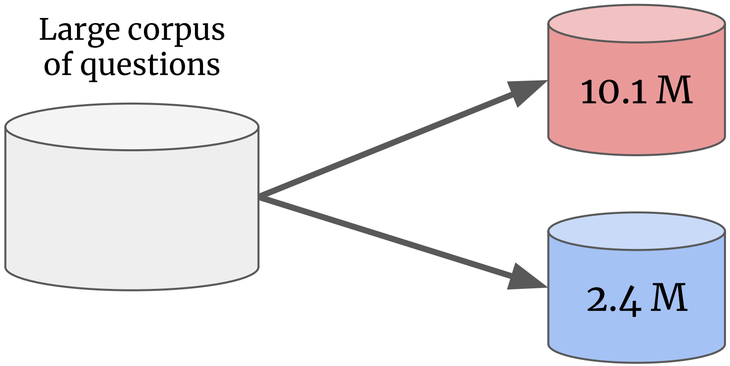
Collected from Common Crawl
using a simple Wh* + ? heuristic

Question corpus creation



Question corpus creation

Large corpus of questions



Large corpus of **single-hop** questions

Large corpus of **multi-hop** questions

Collected from Common Crawl using a simple Wh* + ? heuristic

Using a classifier trained with:
- SQuAD (single-hop)
- Hotpot QA [1] (multi-hop)

[1] Yang et al., EMNLP 2018

Retrieval-based decomposition

$$(s_1, s_2, \ldots, s_N) = d' = \underset{d' \subset S}{\operatorname{argmax}} \sum_{s_i \in d'} f(q, s_i) - \sum_{s_i, s_j \in d', i \neq j} f(s_i, s_j)$$

d' pseudo-decomposition
q question
s_i candidate
f metric (cosine similarity)

Retrieval-based decomposition

$$\underbrace{(s_1, s_2, \ldots, s_N)}_{\text{pseudo-decompositions}} = d' = \operatorname*{argmax}_{d' \subset S} \sum_{s_i \in d'} f(q, s_i) - \sum_{s_i, s_j \in d', i \neq j} f(s_i, s_j)$$

d' pseudo-decomposition
q question
s_i candidate
f metric (cosine similarity)

Retrieval-based decomposition

$$(s_1, s_2, \ldots, s_N) = d' = \underset{d' \subset S}{\operatorname{argmax}} \underbrace{\sum_{s_i \in d'} f(q, s_i)}_{\text{Maximize similarity between questions and retrieved decompositions}} - \sum_{s_i, s_j \in d', i \neq j} f(s_i, s_j)$$

d' pseudo-decomposition
q question
s_i candidate
f metric (cosine similarity)

SAPIENZA NLP

Retrieval-based decomposition

$$(s_1, s_2, \dots, s_N) = d' = \underset{d' \subset S}{\operatorname{argmax}} \sum_{s_i \in d'} f(q, s_i) - \underbrace{\sum_{s_i, s_j \in d', i \neq j} f(s_i, s_j)}_{\text{Minimize similarity between retrieved decompositions}}$$

d' pseudo-decomposition
q question
s_i candidate
f metric (cosine similarity)

Retrieval-based decomposition

$$(s_1, s_2, \ldots, s_N) = d' = \operatorname*{argmax}_{d' \subset S} \sum_{s_i \in d'} f(q, s_i) - \sum_{s_i, s_j \in d', i \neq j} f(s_i, s_j)$$

d' pseudo-decomposition
q question
s_i candidate
f metric (cosine similarity)

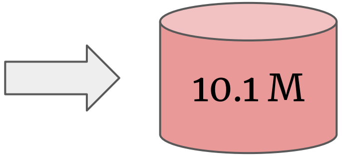
What profession do H. L. Mencken and Albert Camus have in common?

Retrieval-based decomposition

$$(s_1, s_2, \ldots, s_N) = d' = \operatorname*{argmax}_{d' \subset S} \sum_{s_i \in d'} f(q, s_i) - \sum_{s_i, s_j \in d', i \neq j} f(s_i, s_j)$$

d' pseudo-decomposition
q question
s_i candidate
f metric (cosine similarity)

```
What profession do H. L. Mencken
and Albert Camus have in common?
```



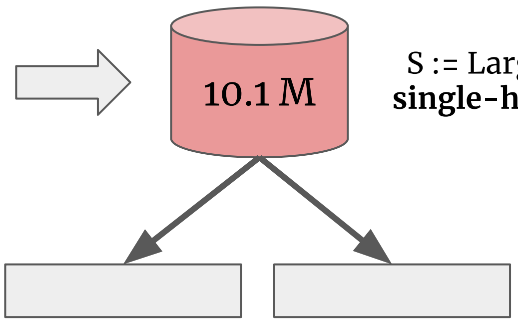
S := Large corpus of **single-hop** questions

Retrieval-based decomposition

$$(s_1, s_2, \ldots, s_N) = d' = \underset{d' \subset S}{\operatorname{argmax}} \sum_{s_i \in d'} f(q, s_i) - \sum_{s_i, s_j \in d', i \neq j} f(s_i, s_j)$$

d' pseudo-decomposition
q question
s_i candidate
f metric (cosine similarity)

`What profession do H. L. Mencken and Albert Camus have in common?`



S := Large corpus of **single-hop** questions

N = 2

$d' = \{s_1^*, s_2^*\}$

Retrieval-based decomposition

$$(s_1^*, s_2^*) = \underset{\{s_1, s_2\} \in S}{\operatorname{argmax}} \left[\hat{\mathbf{v}}_q^\top \hat{\mathbf{v}}_{s_1} + \hat{\mathbf{v}}_q^\top \hat{\mathbf{v}}_{s_2} - \hat{\mathbf{v}}_{s_1}^\top \hat{\mathbf{v}}_{s_2} \right]$$

$\hat{\mathbf{v}}$ unit vector

SAPIENZA NLP

Retrieval-based decomposition

$$(s_1^*, s_2^*) = \underset{\{s_1, s_2\} \in S}{\operatorname{argmax}} \left[\hat{\mathbf{v}}_q^\top \hat{\mathbf{v}}_{s_1} + \hat{\mathbf{v}}_q^\top \hat{\mathbf{v}}_{s_2} - \hat{\mathbf{v}}_{s_1}^\top \hat{\mathbf{v}}_{s_2} \right]$$

$\hat{\mathbf{v}}$ unit vector

Since these comparisons are $O(|S|^2)$ and $|S|>10M$

Retrieval-based decomposition

$$(s_1^*, s_2^*) = \underset{\{s_1,s_2\}\in S}{\mathrm{argmax}} \left[\hat{\mathbf{v}}_q^\top \hat{\mathbf{v}}_{s_1} + \hat{\mathbf{v}}_q^\top \hat{\mathbf{v}}_{s_2} - \hat{\mathbf{v}}_{s_1}^\top \hat{\mathbf{v}}_{s_2}\right]$$

$\hat{\mathbf{v}}$ unit vector

Since these comparisons are $O(|S|^2)$ and $|S|>10M$

$$S' = \mathrm{topK}_{\{s\in S\}} \left[\hat{\mathbf{v}}_q^\top \hat{\mathbf{v}}_s\right]$$

SAPIENZA NLP

Multi-hop to single-hop

$q \Rightarrow d'$

[1] Lample and Conneau, 2019

Multi-hop to single-hop

$q \Rightarrow d'$

- **No learning**: directly use d' = $[s_1, s_2]$ as sub-questions

[1] Lample and Conneau, 2019

Multi-hop to single-hop

$$q \Rightarrow d'$$

- **No learning**: directly use d' = [s_1, s_2] as sub-questions
- **Seq2Seq**: maximize P(d'|q)

[1] Lample and Conneau, 2019

Multi-hop to single-hop

$$q \Rightarrow d'$$

- **No learning**: directly use Use d' = [s_1, s_2] as sub-questions
- **Seq2Seq**: maximize P(d'|q)
- **Unsup. Seq2Seq**: learn mapping q → d, similar to XLM [1], through:
  - denoising,
  - back-translation.

[1] Lample and Conneau, 2019

Multi-hop to single-hop

$$q \Rightarrow d'$$

- **No learning**: directly use Use d’ = [s_1, s_2] as sub-questions
- **Seq2Seq**: maximize P(d’|q)
- **Unsup. Seq2Seq**: learn mapping q → d, similar to XLM [1], through:
  - denoising,
  - back-translation.

Note: in the first two methods, entities in [s_1, s_2] are replaced with entities from q

[1] Lample and Conneau, 2019

Examples

Q1: Are both Coldplay and Pierre Bouvier from the same country?

SQ$_1$: Where are Coldplay and Coldplay from?

∟ Coldplay are a <u>British</u> rock band formed in 1996 by lead vocalist and keyboardist Chris Martin and lead guitarist Jonny Buckland at University College London (UCL).

SQ$_2$: What country is Pierre Bouvier from?

∟ Pierre Charles Bouvier (born 9 May 1979) is a <u>Canadian</u> singer, songwriter, musician, composer and actor who is best known as the lead singer and guitarist of the rock band Simple Plan.

$\hat{\mathbf{A}}$: No

Examples

Q2: How many copies of Roald Dahl's variation on a popular anecdote sold?

SQ$_1$: How many copies of Roald Dahl's?

∟ His books have sold more than 250 million copies worldwide.

SQ$_2$ What is the name of the variation on a popular anecdote?

∟ "Mrs. Bixby and the Colonel's Coat" is a short story by Roald Dahl that first appeared in the 1959 issue of Nugget.

Â: more than 250 million

SAPIENZA NLP

Examples

Q3: Who is older, Annie Morton or Terry Richardson?

SQ$_1$: Who is Annie Morton?

⌞ Annie Morton (born October 8, 1970) is an American model born in Pennsylvania.

SQ$_2$: When was Terry Richardson born?

⌞ Kenton Terry Richardson (born 26 July 1999) is an English professional footballer who plays as a defender for League Two side Hartlepool United.

$\hat{\mathbf{A}}$: Annie Morton

Results on HotpotQA (with/without decomp.)

Q-Type	Using Decomps. ✗	Using Decomps. ✓
Bridge	$80.1_{\pm.2}$	$\mathbf{81.7}_{\pm.4}$
Comp.	$73.8_{\pm.4}$	$\mathbf{80.1}_{\pm.3}$
Inters.	$79.4_{\pm.6}$	$\mathbf{82.3}_{\pm.5}$
1-hop	$73.9_{\pm.6}$	$\mathbf{76.9}_{\pm.6}$

SAPIENZA NLP

Comparison

Decomp. Method	Pseudo-Decomps.	HotpotQA Dev F1 Orig	Multi	OOD
✗	✗ (1hop)	66.7	63.7	66.5
✗	✗ (Baseline)	77.0$_{\pm .2}$	65.2$_{\pm .2}$	67.1$_{\pm .5}$
PseudoD	Random	78.4$_{\pm .2}$	70.9$_{\pm .2}$	70.7$_{\pm .4}$
	FastText	78.9$_{\pm .2}$	72.4$_{\pm .1}$	72.0$_{\pm .1}$
Seq2Seq	Random	77.7$_{\pm .2}$	69.4$_{\pm .3}$	70.0$_{\pm .7}$
	FastText	78.9$_{\pm .2}$	73.1$_{\pm .2}$	73.0$_{\pm .3}$
ONUS	Random	79.8$_{\pm .1}$	76.0$_{\pm .2}$	76.5$_{\pm .2}$
	FastText	**80.1**$_{\pm .2}$	**76.2**$_{\pm .1}$	**77.1**$_{\pm .1}$
DecompRC*		79.8$_{\pm .2}$	76.3$_{\pm .4}$	77.7$_{\pm .2}$
SAE (Tu et al., 2020) †		80.2	61.1	62.6
HGN (Fang et al., 2019) †		82.2	78.9‡	76.1‡

Comparison

Decomp. Method	Pseudo-Decomps.	HotpotQA Dev F1 Orig	Multi	OOD
✗	✗ (1hop)	66.7	63.7	66.5
✗	✗ (Baseline)	77.0$_{\pm.2}$	65.2$_{\pm.2}$	67.1$_{\pm.5}$
PseudoD	Random	78.4$_{\pm.2}$	70.9$_{\pm.2}$	70.7$_{\pm.4}$
	FastText	78.9$_{\pm.2}$	72.4$_{\pm.1}$	72.0$_{\pm.1}$
Seq2Seq	Random	77.7$_{\pm.2}$	69.4$_{\pm.3}$	70.0$_{\pm.7}$
	FastText	78.9$_{\pm.2}$	73.1$_{\pm.2}$	73.0$_{\pm.3}$
ONUS	Random	79.8$_{\pm.1}$	76.0$_{\pm.2}$	76.5$_{\pm.2}$
	FastText	**80.1**$_{\pm.2}$	**76.2**$_{\pm.1}$	**77.1**$_{\pm.1}$
DecompRC*		79.8$_{\pm.2}$	76.3$_{\pm.4}$	77.7$_{\pm.2}$
SAE (Tu et al., 2020) †		80.2	61.1	62.6
HGN (Fang et al., 2019) †		82.2	78.9‡	76.1‡

Baselines

Comparison

Decomp. Method	Pseudo-Decomps.	HotpotQA Dev F1 Orig	Multi	OOD
✗	✗ (1hop)	66.7	63.7	66.5
✗	✗ (Baseline)	$77.0_{\pm.2}$	$65.2_{\pm.2}$	$67.1_{\pm.5}$
PseudoD	Random	$78.4_{\pm.2}$	$70.9_{\pm.2}$	$70.7_{\pm.4}$
	FastText	$78.9_{\pm.2}$	$72.4_{\pm.1}$	$72.0_{\pm.1}$
Seq2Seq	Random	$77.7_{\pm.2}$	$69.4_{\pm.3}$	$70.0_{\pm.7}$
	FastText	$78.9_{\pm.2}$	$73.1_{\pm.2}$	$73.0_{\pm.3}$
ONUS	Random	$79.8_{\pm.1}$	$76.0_{\pm.2}$	$76.5_{\pm.2}$
	FastText	**80.1**$_{\pm.2}$	**76.2**$_{\pm.1}$	**77.1**$_{\pm.1}$
DecompRC*		$79.8_{\pm.2}$	$76.3_{\pm.4}$	$77.7_{\pm.2}$
SAE (Tu et al., 2020) †		80.2	61.1	62.6
HGN (Fang et al., 2019) †		82.2	78.9‡	76.1‡

Baselines

Related works
(using supervision)

SAPIENZA NLP

Conclusion

Conclusion

We have seen two impactful unsupervised approaches for QA:

- creation of synthetic training data,
- decomposition of hard questions into simpler ones.

Conclusion

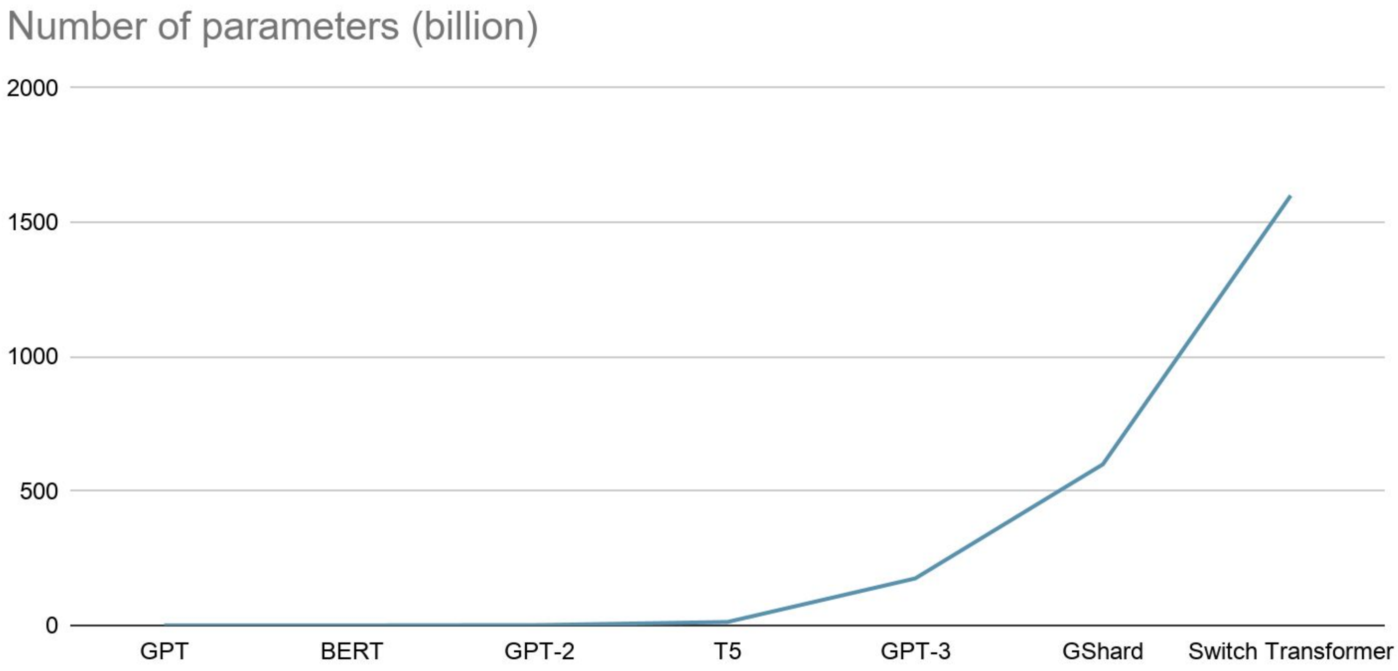
We have seen two impactful unsupervised approaches for QA:

- creation of synthetic training data,
- decomposition of hard questions into simpler ones.

Advantages:

- scalable,
- can be adapted to new domains, depending on the need.

Recent trends in Deep Learning architectures



Conclusion

Do we really need labeled data?

Conclusion

Do we really need labeled data?

Yes.

Thank you for your attention!

Come visit us at http://nlp.uniroma1.it/

SAPIENZA NLP

SAPIENZA UNIVERSITÀ DI ROMA