On Transformers, Distributional Models and Typicality of Argument Structures

Emmanuele Chersoni

The Hong Kong Polytechnic University

Outline of the Talk

- Challenging Transformers with Generalized Event Knowledge
- The Problem of Logical Metonymy Interpretation (Appendix)

Event-Based Priming in Sentence Processing

- **Generalized Event Knowledge (GEK)** in human sentence processing
- knowledge about events and typical participants → used by humans to speed up the comprehension process
  - verbs activate typical arguments and *vice versa* (McRae et al., 1998; 2005)
  - nouns activate typical co-arguments (Hare et al., 2009)

Event-Based Priming in Sentence Processing

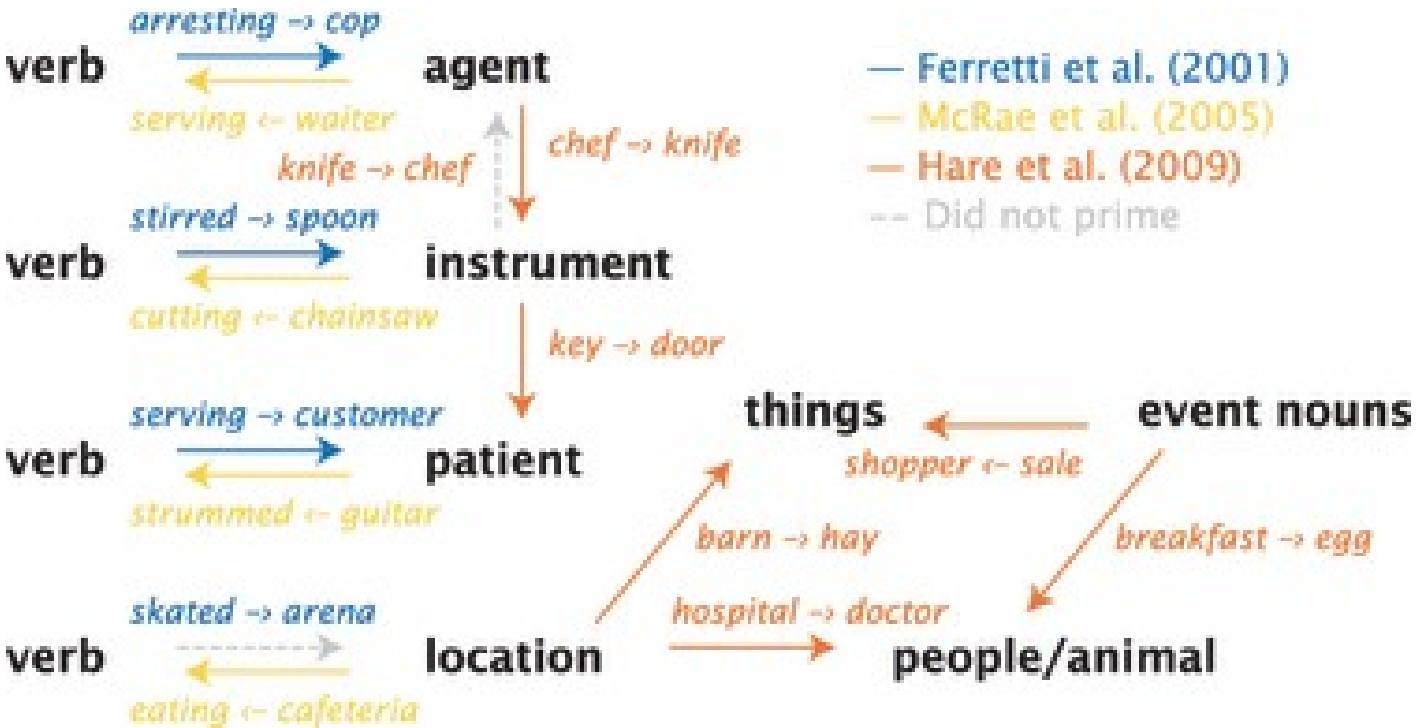

Figure: Scheme of event-based priming findings, as summarized by McRae and Matsuki, 2009. Notice that, in the standard priming setting (naming study with SOA), no priming was found from verbs to typical locations.

Event-Based Priming in Sentence Processing

- processing facilitation in sentence comprehension
  - typical argument combinations i) are read faster in self-paced reading and elicit smaller N400 amplitudes (Bicknell et al., 2010), ii) lead to shorter fixations in eye-tracking (Matsuki et al., 2011)
    - *The secretary is addressing the boss/the speaker is addressing the audience* (TYP)
    - *the speaker is addressing the boss/the secretary is addressing the audience* (NON TYP)
- words in the mental lexicon arranged as a *network of mutual expectations* (McRae and Matsuki, 2009)
- thematic fit $\rightarrow$ "degree of typicality" of an argument filler of a given verb role (McRae et al., 1998)

Event-Based Priming in Sentence Processing

- processing facilitation in sentence comprehension
  - typical argument combinations i) are read faster in self-paced reading and elicit smaller N400 amplitudes (Bicknell et al., 2010), ii) lead to shorter fixations in eye-tracking (Matsuki et al., 2011)
    - *The secretary is addressing the boss/the speaker is addressing the audience* (TYP)
    - *the speaker is addressing the boss/the secretary is addressing the audience* (NON TYP)
- words in the mental lexicon arranged as a *network of mutual expectations* (McRae and Matsuki, 2009)
- thematic fit $\rightarrow$ ”degree of typicality” of an argument filler of a given verb role (McRae et al., 1998)

Distributional Models of the GEK

- psycholinguistic research made available human thematic fit judgements
- vector-based cosine similarity as a proxy of human typicality judgements (Baroni and Lenci, 2010; Lenci, 2011; Greenberg et al., 2015; Sayeed et al., 2016; Santus et al., 2017; Chersoni et al., 2019)
- main idea: clusters of typical filler vectors (*filler prototypes*) to model expectations about upcoming roles

Distributional Models of the GEK

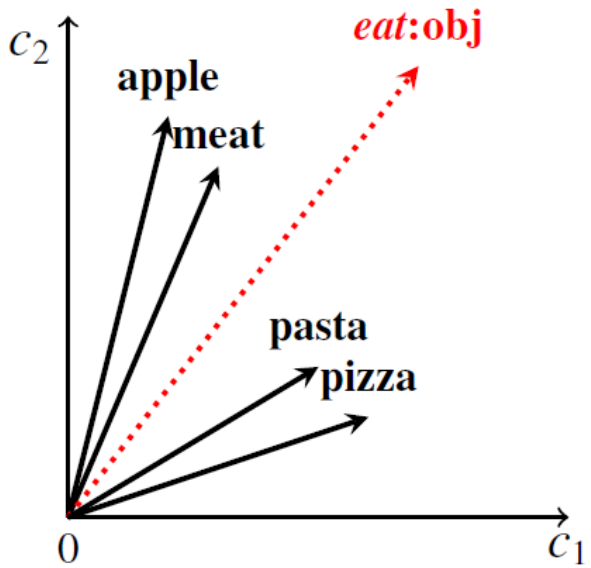

Figure: Filler prototype construction for the object of *to eat* by clustering the vectors of typical fillers.

Transformer Models (of the GEK?)

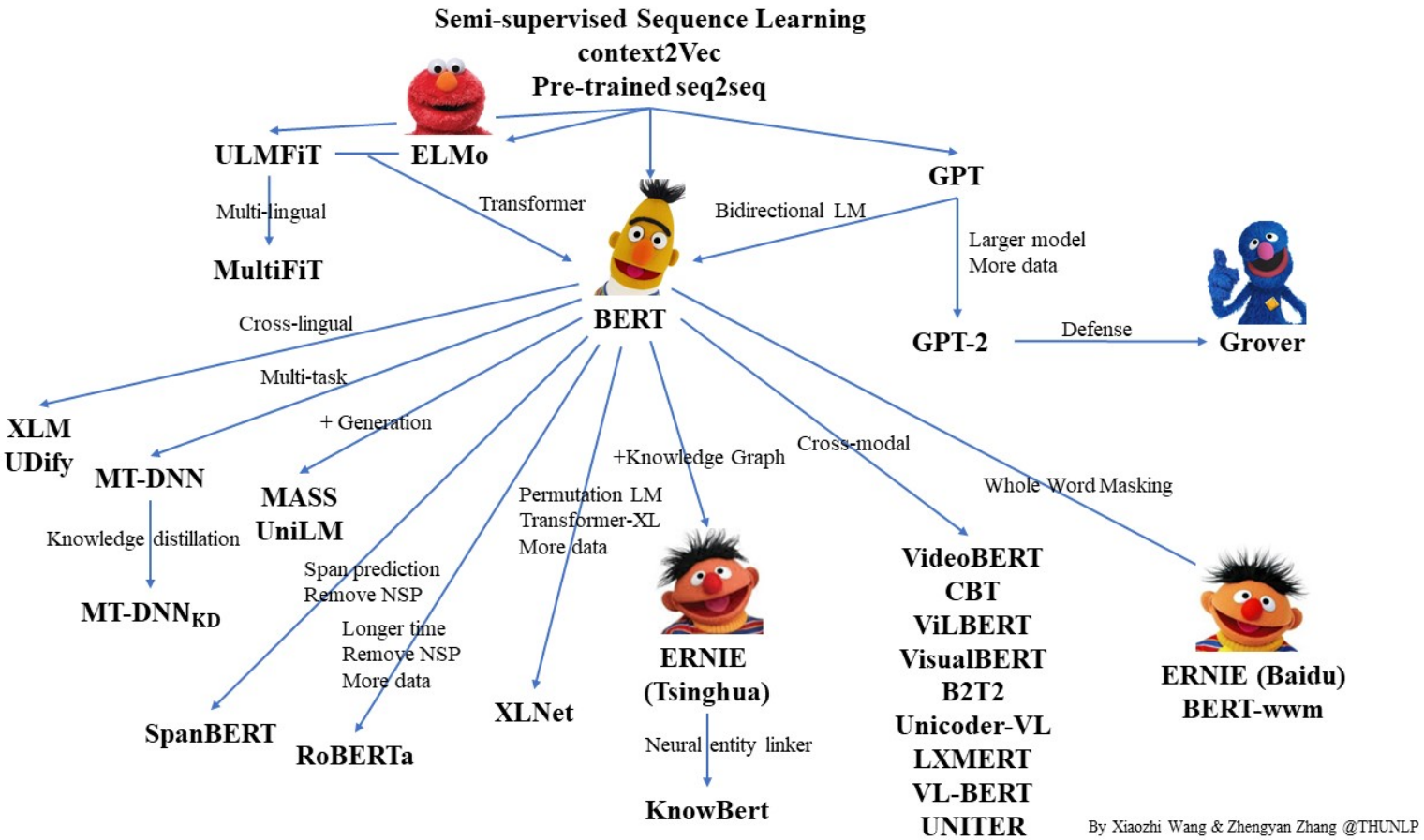

Figure: Scheme summarizing the evolution of Transformer models. From the Github page of Zhengyan Zhang and Xiaozhi Wang: `https://github.com/thunlp/PLMpapers`.

Transformer Models (of the GEK?)

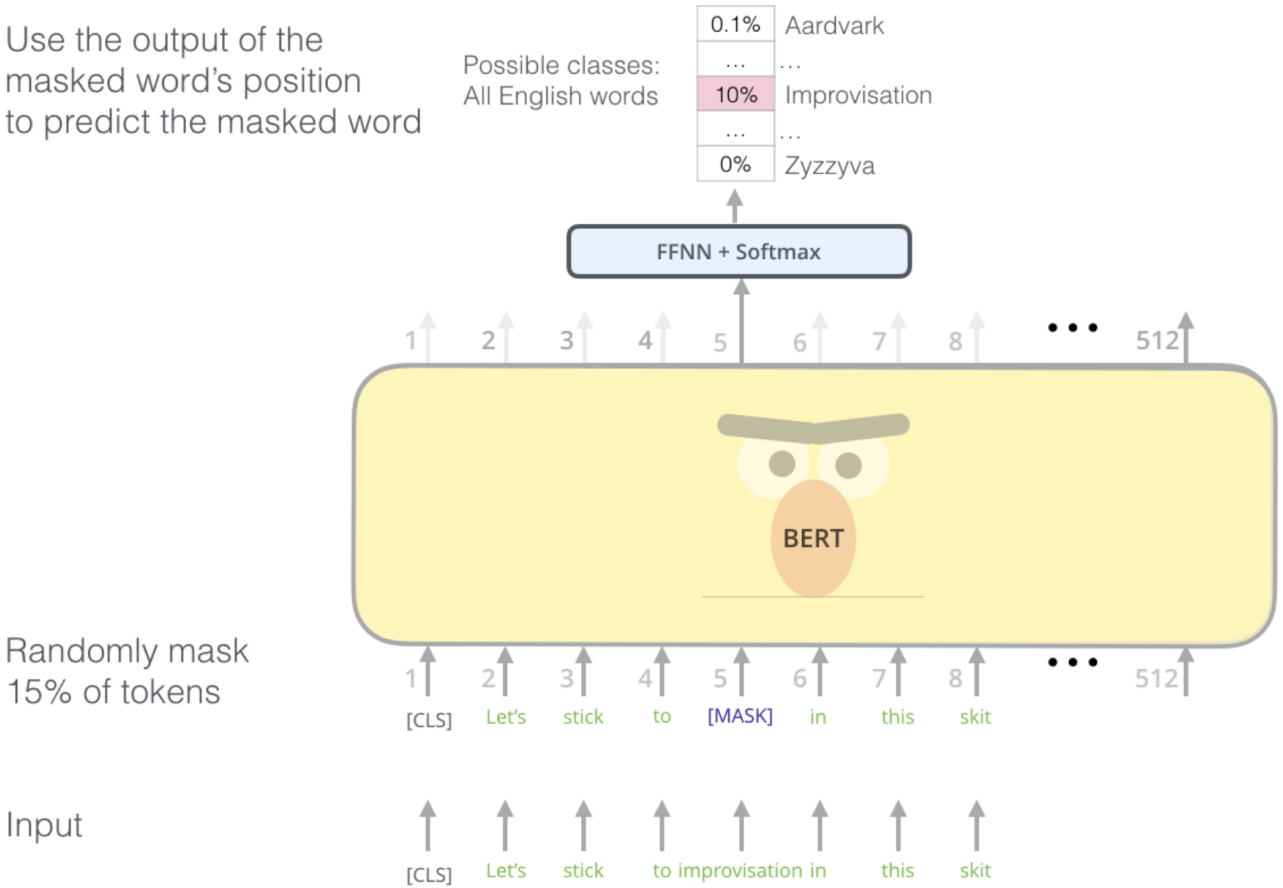

Figure: Image taken from Jay Alammar's blog: `http://jalammar.github.io/illustrated-bert/`.

Transformer Models (of the GEK?)

- Transformer architectures are trained on guessing words given a context (*masked language modeling*)
  - despite being trained only on text, they have been shown to encode a lot of relational and factual knowledge (Petroni et al., 2019)
- GEK comes from both first-hand experience of the world and from linguistic experience (McRae and Matsuki, 2009; Chersoni et al., 2019)
- **research question**: do Transformers encode (at least partially) GEK?

Transformer Models (of the GEK?)

- Transformer architectures are trained on guessing words given a context (*masked language modeling*)
  - despite being trained only on text, they have been shown to encode a lot of relational and factual knowledge (Petroni et al., 2019)
- GEK comes from both first-hand experience of the world and from linguistic experience (McRae and Matsuki, 2009; Chersoni et al., 2019)
- **research question**: do Transformers encode (at least partially) GEK?

Transformer Models (of the GEK?)

- Transformer architectures are trained on guessing words given a context (*masked language modeling*)
  - despite being trained only on text, they have been shown to encode a lot of relational and factual knowledge (Petroni et al., 2019)
- GEK comes from both first-hand experience of the world and from linguistic experience (McRae and Matsuki, 2009; Chersoni et al., 2019)
- **research question**: do Transformers encode (at least partially) GEK?

The DTFit Dataset

- DTFit (Vassallo et al., 2018): a collection of pairs of argument tuples for dynamic thematic fit estimation
- tuples differ for only one target element, which changes the event typicality
- divided in different subsets, according to the role of the target filler

Role	Tuple	Typical	Atypical
Agent	__ mix paint	painter	cook
Patient	tailor sew __	dress	wound
Instrument	cook clean fish __	knife	sponge
Time	cat chase bird __	hunting	marriage
Location	sailor mop deck __	boat	theatre

Table: Examples of tuples from DTFit.

The DTFit Dataset

- collection of typicality ratings from native English speakers (1: very atypical, 7: very typical)
- knowledge about professions, but also everyday life situations
- notice that the fillers generally have a good fit with the verb in isolation, but differ in typicality in the general scenario
  - compare *mix-soup* with *the painter mixed the soup*

The DTFit Dataset

- collection of typicality ratings from native English speakers (1: very atypical, 7: very typical)
- knowledge about professions, but also everyday life situations
- notice that the fillers generally have a good fit with the verb in isolation, but differ in typicality in the general scenario
  - compare *mix-soup* with *the painter mixed the soup*

The DTFit Dataset

- Task for the models: predict a target argument, given the other ones
  - *The painter mixes the .... ?*
- Evaluation metrics:
  - *Spearman correlation* between the output scores and the human-elicited judgements
  - (only for agent-patient tuples) *Accuracy* in assigning a higher thematic fit score to typical tuples

Baseline Model: SDM

- Structured Distributional Model (SDM; Chersoni et al., 2019)
  - distributional model relying on a graph of verb-argument relations extracted from a 2018 Wikipedia dump
  - estimation of filler prototypes takes context into account: graph edges reproduce reciprocal argument activations
  - as in traditional DSM, the thematic fit of a filler is similarity with the prototype (Baroni and Lenci, 2010)

Distributional Models of the GEK

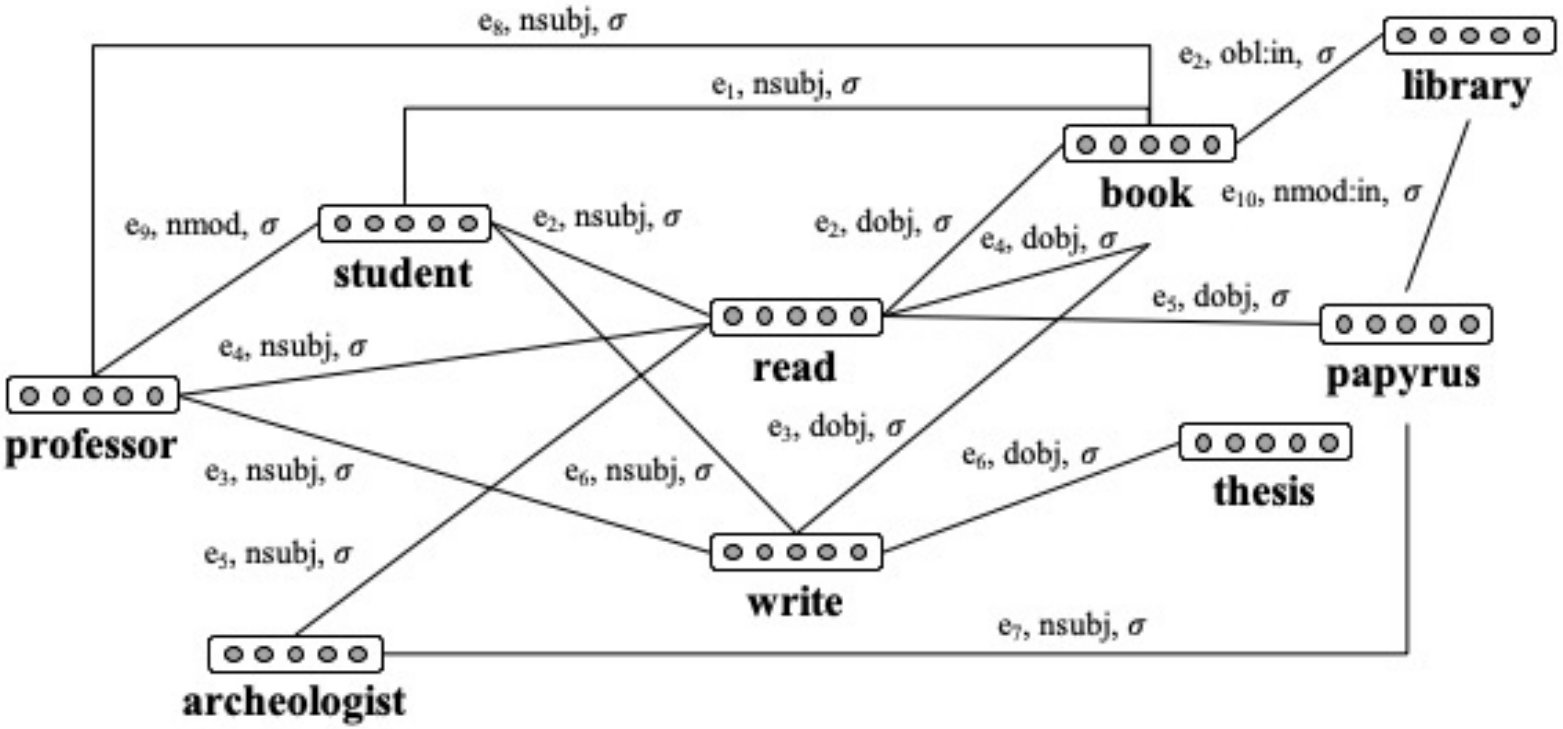

Figure: Illustration of the GEK as a network of mutual expectations from Chersoni et al. (2019).

Transformer Models

- we adopted some of the most popular Transformer models: BERT (both Base and Large; Devlin et al., 2019), GPT-2 (Radford et al., 2019) and RoBERTa (Liu et al., 2019)
- for each dataset tuple, we append definite articles to derive simple sentences and mask the target words
  - *tailor sew dress* $\rightarrow$ The tailor sewed the dress $\rightarrow$ The tailor sewed the [MASK]
- the thematic fit score is the probability score assigned to the target filler

Results

	Coverage	SDM	BERT-base	BERT-large	ROBERTA-large	GPT-2
Agent$_{\text{DTFit}}$	105/134	0.58	0.46	0.53	**0.64**	-
Patient$_{\text{DTFit}}$	323/402	0.62	0.59	**0.64**	**0.64**	0.63
Instrument$_{\text{DTFit}}$	31/100	**0.58**	0.52	0.53	0.5	0.5
Time$_{\text{DTFit}}$	89/100	0.58	0.63	0.64	**0.66**	**0.66**
Location$_{\text{DTFit}}$	115/150	0.65	0.72	0.71	0.73	**0.74**

Table: Spearman Correlation for the DTFit datasets (TLMs and SDM differences are significant only for the Locations subset). Notice that, given the different coverage of the models, we had to compute the correlations on the subset of the common tuples.

Results

	DTFit		Wang2018	
	Agent	Patient	Agent	Patient
SDM	**.89**	**.91**	.65	.66
BERT-base	.77	.85	.76	.63
BERT-large	.83	.89	**.77**	.65
ROBERTA-large	**.89**	**.91**	.76	**.73**
GPT-2 medium	-	.90	-	.64

Table: Accuracy in the binary classification task for DTFit (agent and patient roles) and Wang2018 datasets.

Results

- First takeaways:
  - generally robust correlation/accuracy scores for all models
  - no significant differences between SDM and Transformer models, although Transformer are trained on much bigger corpora

Results

Tuple	Expected	Preferred
mason mix __	cement	soup
climber climb __	rock	staircase
blacksmith pour __	metal	wine
chemist pour __	compound	juice
cat drink __	milk	coffee

Table: Examples of errors (BERT-base model, $\text{Patient}_{\text{DTFit}}$). Notice that, in many cases, the patient is plausible but the composition with the agent is not taken into account.

Diagnostic Tests

1. do Transformers just prefer frequent collocations?
2. do Transformers recognize semantic plausibility?
3. do Transformers extend thematic fit judgements to low-frequency synonyms/hyponyms?
4. are Transformers influenced by the syntactic structure of the sentences?

Diagnostic Tests

Question 1: do Transformers just prefer frequent collocations?

- they might predict frequent complements of the verbs, irrespective of coherence with other arguments
- we create a small dataset of 31 pairs from $\textbf{Patient}_{\text{DTFit}}$, replacing the atypical patient with a noun with high statistical association with the verb, but not coherent with the general context
  - *The terrorist released the **album**.*
  - *The soldier heard the **case**.*
- We tested BERT-Base, BERT-Large and SDM on the binary classification

Diagnostic Tests

Question 1: do Transformers just prefer frequent collocations?

- they might predict frequent complements of the verbs, irrespective of coherence with other arguments
- we create a small dataset of 31 pairs from $\textbf{Patient}_{\text{DTFit}}$, replacing the atypical patient with a noun with high statistical association with the verb, but not coherent with the general context
  - *The terrorist released the* ***album***.
  - *The soldier heard the* ***case***.
- We tested BERT-Base, BERT-Large and SDM on the binary classification

Diagnostic Tests

Question 1: do Transformers just prefer frequent collocations?

- BERT-Base and Large perform quite well: only 9 and 6 wrong pairs
- atypical events are picked only when the contrast between atypical fillers and expectations is not so evident
  - *The smuggler sold the **property*** is preferred to *The smuggler sold the **weapon***.
- SDM got 14 wrong pairs: sparsity problem (no co-occurrences between agents and patients)

Diagnostic Tests

Question 2: do Transformers recognize semantic plausibility?

- plausibility is more challenging for computational models than typicality: what is typical is also frequent, but what is plausible not necessarily
- Wang2018 dataset (Wang et al., 2018): physically plausible vs. physically implausible triplets, both unattested in corpora
  - The student climbed the ship (plausible)
  - The student climbed the water (implausible)
- we built two datasets of 222 and 394 triple pairs, respectively, differing either for the agent or for the patient filler, and tested the models in binary classification

Diagnostic Tests

Question 2: do Transformers recognize semantic plausibility?

- plausibility is more challenging for computational models than typicality: what is typical is also frequent, but what is plausible not necessarily
- Wang2018 dataset (Wang et al., 2018): physically plausible vs. physically implausible triplets, both unattested in corpora
  - The student climbed the ship (plausible)
  - The student climbed the water (implausible)
- we built two datasets of 222 and 394 triple pairs, respectively, differing either for the agent or for the patient filler, and tested the models in binary classification

Diagnostic Tests

Question 2: do Transformers recognize semantic plausibility?

- plausibility is more challenging for computational models than typicality: what is typical is also frequent, but what is plausible not necessarily
- Wang2018 dataset (Wang et al., 2018): physically plausible vs. physically implausible triplets, both unattested in corpora
  - The student climbed the ship (plausible)
  - The student climbed the water (implausible)
- we built two datasets of 222 and 394 triple pairs, respectively, differing either for the agent or for the patient filler, and tested the models in binary classification

Diagnostic Tests

Question 2: do Transformers recognize semantic plausibility?

	DTFit		**Wang2018**	
	Agent	Patient	Agent	Patient
SDM	**.89**	**.91**	.65	.66
BERT-base	.77	.85	.76	.63
BERT-large	.83	.89	**.77**	.65
ROBERTA-large	**.89**	**.91**	.76	**.73**
GPT-2 medium	-	.90	-	.64

Table: Accuracy in the binary classification task for DTFit (agent and patient roles) and Wang2018 datasets.

Diagnostic Tests

- plausibility is more difficult to model than typicality when both plausible and implausible events are rare or unattested in the training data
- task generally tackled by injecting some external knowledge in the vector representations (Wang et al., 2018)

Diagnostic Tests

- plausibility is more difficult to model than typicality when both plausible and implausible events are rare or unattested in the training data
- task generally tackled by injecting some external knowledge in the vector representations (Wang et al., 2018)

Diagnostic Tests

Question 3: do Transformers extend thematic fit judgements to low-frequency synonyms/hyponyms?

- test on the generalization capability of the models: if a filler fits well in a role, a rare synonym/hyponym should be equally a good filler
- 39 pairs from $\textbf{Patient}_{\text{DTFit}}$ in which we replaced the filler in the typical condition with a low frequency synonym/hyponym from WordNet (Fellbaum, 1998)
  - *The waitress cleared the* ***restaurant*** $\rightarrow$ ***tavern***
  - *The veterinarian examined the* ***dog*** $\rightarrow$ ***puppy***
- BERT Base and Large and SDM were tested again in binary classification

Diagnostic Tests

Question 3: do Transformers extend thematic fit judgements to low-frequency synonyms/hyponyms?

- the proposed variations pose serious difficulties to the models
  - BERT-Large and SDM barely above random guessing (53% of accuracy)
  - BERT-Base well below it (37%)
- weak generalization capacity from input text

Diagnostic Tests

Question 4: are Transformers influenced by the syntactic structure of the sentences?

- check if the predictions are influenced by recurrent word order patterns
- we created two different versions of the DTFit dataset: one with **wh-interrogative** and one with **cleft sentences**
  - wh-interrogative: *The actor won the award* $\rightarrow$ *Which award did the actor win?*
  - cleft sentences: *The actor won the award* $\rightarrow$ *It was the award that the actor won*
- after masking the target word for the subset, we tested the Spearman correlation scores of RoBERTa-large

Diagnostic Tests

Question 4: are Transformers influenced by the syntactic structure of the sentences?

- check if the predictions are influenced by recurrent word order patterns
- we created two different versions of the DTFit dataset: one with **wh-interrogative** and one with **cleft sentences**
  - wh-interrogative: *The actor won the award* $\rightarrow$ *Which award did the actor win?*
  - cleft sentences: *The actor won the award* $\rightarrow$ *It was the award that the actor won*
- after masking the target word for the subset, we tested the Spearman correlation scores of RoBERTa-large

Diagnostic Tests

Question 4: are Transformers influenced by the syntactic structure of the sentences?

	transitive	cleft	wh-interrogative
Agent$_{\text{DTFit}}$	0.64	-0.13	0.62
Patient$_{\text{DTFit}}$	0.64	0.26	0.51
Instrument$_{\text{DTFit}}$	0.50	0.10	0.60
Time$_{\text{DTFit}}$	0.66	0.33	0.64
Location$_{\text{DTFit}}$	0.73	0.67	0.73

Table: Spearman Correlation for DTFit datasets using RoBERTa-large and input sentences with different word orders.

Diagnostic Tests

Question 4: are Transformers influenced by the syntactic structure of the sentences?

- although the wh-interrogative does not affect too much the scores, cleft structures cause a huge drop
- word predictions in the new construction are more dependent on collocates
  - e.g. *It was with the [MASK] that the guard opened the door*, correctly predicted *key* in transitive setting
  - RoBERTa predicts the following fillers: *gun*, *crowd*, *sword* and then *key*
  - probably focusing on *guard* collocates and ignoring the general context

Diagnostic Tests

Question 4: are Transformers influenced by the syntactic structure of the sentences?

- although the wh-interrogative does not affect too much the scores, cleft structures cause a huge drop
- word predictions in the new construction are more dependent on collocates
  - e.g. *It was with the [MASK] that the guard opened the door*, correctly predicted *key* in transitive setting
  - RoBERTa predicts the following fillers: *gun*, *crowd*, *sword* and then *key*
  - probably focusing on *guard* collocates and ignoring the general context

Conclusion

- Transformers and SDM show comparable performance on GEK-related tasks
- small diagnostic tests to investigate the model performance
  - models still depend a lot on what they observe during training, with limited generalizing ability (a problem shared with DSMs)
  - heavy reliance on frequency and locality of associations

Conclusion

- Reference: Pedinotti P., Rambelli G., Chersoni E., Santus E., Lenci A., Blache P. (2021). Did the cat drink the coffee? Challenging Transformers with Generalized Event Knowledge. Proceedings of StarSEM.

The problem of logical metonymy

- *Logical metonymy* → type clash between an event-selecting verb and an entity-denoting noun, interpreted via the retrieval of a hidden event
  - *The editor finished the article* → READING
- challenging for traditional theories of compositionality (Jackendoff, 1990; Asher, 2015; Pustejovsky and Batiukova, 2019)
- psycholinguistic studies reporting extra processing costs for such constructions (McElree et al., 2001; Traxler et al., 2002)

The problem of logical metonymy

- *Logical metonymy* → type clash between an event-selecting verb and an entity-denoting noun, interpreted via the retrieval of a hidden event
  - *The editor finished the article* → READING
- challenging for traditional theories of compositionality (Jackendoff, 1990; Asher, 2015; Pustejovsky and Batiukova, 2019)
- psycholinguistic studies reporting extra processing costs for such constructions (McElree et al., 2001; Traxler et al., 2002)

The problem of logical metonymy

- *Logical metonymy* $\rightarrow$ type clash between an event-selecting verb and an entity-denoting noun, interpreted via the retrieval of a hidden event
  - *The editor finished the article* $\rightarrow$ READING
- challenging for traditional theories of compositionality (Jackendoff, 1990; Asher, 2015; Pustejovsky and Batiukova, 2019)
- psycholinguistic studies reporting extra processing costs for such constructions (McElree et al., 2001; Traxler et al., 2002)

The problem of logical metonymy

- logical metonymies have been explained in the Generalized Event Knowledge framework (McRae and Matsuki, 2009)
- speakers access event knowledge during sentence processing, inferring the most likely event given the contextual cues (Zarcone et al., 2014)

The problem of logical metonymy

Logical metonymy interpretation in NLP research:

- probabilistic methods (Lapata and Lascarides, 2003; Zarcone et al., 2012)
- distributional models (Zarcone et al. 2012; 2013; Chersoni et al., 2017)
- what about the newly-introduced transformer models (BERT and co.) ?

Goal of the task: given a metonymic expression, **retrieve the hidden event** corresponding to the interpretation of the metonymic expression

The problem of logical metonymy

Logical metonymy interpretation in NLP research:

- probabilistic methods (Lapata and Lascarides, 2003; Zarcone et al., 2012)
- distributional models (Zarcone et al. 2012; 2013; Chersoni et al., 2017)
- what about the newly-introduced transformer models (BERT and co.) ?

Goal of the task: given a metonymic expression, **retrieve the hidden event** corresponding to the interpretation of the metonymic expression

The problem of logical metonymy

Logical metonymy interpretation in NLP research:

- probabilistic methods (Lapata and Lascarides, 2003; Zarcone et al., 2012)
- distributional models (Zarcone et al. 2012; 2013; Chersoni et al., 2017)
- what about the newly-introduced transformer models (BERT and co.) ?

Goal of the task: given a metonymic expression, **retrieve the hidden event** corresponding to the interpretation of the metonymic expression

Classical benchmarks

McElree et al. (2001) and Traxler et al. (2002)

- 30 sentences obtained from the self-paced reading experiment of McElree et al. and 36 from the eye-tracking experiment of Traxler et al.
  - The author started (*writing*) the book (HIGH TYPICALITY condition)
  - The author started (*reading*) the book (LOW TYPICALITY condition)
- a common finding: in the highly-typical condition for the interpretation of the verb, faster processing times
- the task is again a **binary classification**: a system has to assign a higher score to the most typical one

Classical benchmarks

McElree et al. (2001) and Traxler et al. (2002)

- 30 sentences obtained from the self-paced reading experiment of McElree et al. and 36 from the eye-tracking experiment of Traxler et al.
  - The author started (*writing*) the book (HIGH TYPICALITY condition)
  - The author started (*reading*) the book (LOW TYPICALITY condition)
- a common finding: in the highly-typical condition for the interpretation of the verb, faster processing times
- the task is again a **binary classification**: a system has to assign a higher score to the most typical one

Classical benchmarks

McElree et al. (2001) and Traxler et al. (2002)

- 30 sentences obtained from the self-paced reading experiment of McElree et al. and 36 from the eye-tracking experiment of Traxler et al.
  - The author started (*writing*) the book (HIGH TYPICALITY condition)
  - The author started (*reading*) the book (LOW TYPICALITY condition)
- a common finding: in the highly-typical condition for the interpretation of the verb, faster processing times
- the task is again a **binary classification**: a system has to assign a higher score to the most typical one

Classical benchmarks

Lapata and Lascarides (2003)

- 174 tuples including a metonymic verb, a direct object and possible hidden verbs, together with mean human plausibility ratings
  - start a letter → *writing* 0.47
  - start a letter → *studying* 0.06
- Task: measure the **correlation** between human judgements and model scores

Classical benchmarks

Lapata and Lascarides (2003)

- 174 tuples including a metonymic verb, a direct object and possible hidden verbs, together with mean human plausibility ratings
  - start a letter → *writing* 0.47
  - start a letter → *studying* 0.06
- Task: measure the **correlation** between human judgements and model scores

Classical benchmarks

Lapata and Lascarides (2003)

- 174 tuples including a metonymic verb, a direct object and possible hidden verbs, together with mean human plausibility ratings
  - start a letter → *writing* 0.47
  - start a letter → *studying* 0.06
- Task: measure the **correlation** between human judgements and model scores

Classical benchmarks

Lenci (forthcoming)

- with SVO templates taken from McElree and Traxler datasets, human subjects are asked to produce two plausible verbs for the interpretation
- 69 SVO templates, 4084 collected verbs but...
- ... very sparse data, we retained only the 285 with frequency of at least 3
  - pilot master plane $\rightarrow$ *flying* 0.56
  - pilot master plane $\rightarrow$ *landing* 0.18
- Task: measure the **correlation** between production frequency and model scores

Classical benchmarks

Lenci (forthcoming)

- with SVO templates taken from McElree and Traxler datasets, human subjects are asked to produce two plausible verbs for the interpretation
- 69 SVO templates, 4084 collected verbs but...
- ... very sparse data, we retained only the 285 with frequency of at least 3
  - pilot master plane $\rightarrow$ *flying* 0.56
  - pilot master plane $\rightarrow$ *landing* 0.18
- Task: measure the **correlation** between production frequency and model scores

Classical benchmarks

Lenci (forthcoming)

- with SVO templates taken from McElree and Traxler datasets, human subjects are asked to produce plausible verbs for the interpretation
- 69 SVO templates, 4084 collected verbs but...
- ... very sparse data, we retained only the 285 with frequency of at least 3
  - pilot master plane $\rightarrow$ *flying* 0.56
  - pilot master plane $\rightarrow$ *landing* 0.18
- Task: measure the **correlation** between production frequency and model scores

Classical benchmarks

Lenci (forthcoming)

- with SVO templates taken from McElree and Traxler datasets, human subjects are asked to produce plausible verbs for the interpretation
- 69 SVO templates, 4084 collected verbs but...
- ... very sparse data, we retained only the 285 with frequency of at least 3
  - pilot master plane $\rightarrow$ *flying* 0.56
  - pilot master plane $\rightarrow$ *landing* 0.18
- Task: measure the **correlation** between production frequency and model scores

Computational models

- Structured Distributional Model (SDM, Chersoni et al., 2019)
  - build a prototype vector of a typical event, given the agent and the patient (e.g. prototype of the verb given *author* as agent and *novel* as patient)
- Transformer models: BERT (Devlin et al., 2019), GPT-2 (Radford et al., 2019) and RoBERTa (Liu et al., 2019)
- with Transformers we treat the problem as a covert event prediction task: we add an additional token for the hidden event
  - The architect finishes the house $\rightarrow$ The architect finishes [MASK] the house.

Evaluation Results

Dataset	SDM	BERT	RoBERTa	GPT-2
McEl	0.77	0.70	0.80	0.87
Trax	0.72	0.47	0.72	0.69

Accuracy scores for the binary classification task (Datasets: McEl = McElree et al., 2001; Trax = Traxler et al., 2002)

Dataset	SDM	BERT	RoBERTa	GPT-2
L	0.40	0.37	0.39	0.31
L&L	0.53	0.61	0.73	0.43

Spearman correlation scores (L = Lenci, L&L = Lapata and Lascarides, 2003)

Evaluation Results

- distributional models
  - robust performance by SDM: top score on two datasets, almost always close to RoBERTa
- transformer-based models
  - RoBERTa is the only transformer that performs consistently across all benchmarks, probably thanks also to the gigantic size of the training corpus
  - GPT-2 achieves the top score on McElree et al. (2001), but it disappoints on all other benchmarks

Error Analysis

- consistently mistaken items for the McElree and the Traxler datasets:
  - *The teenager starts the novel.* (McEl)
  - *The teenager begins the novel.* (Trax)
- in all the above cases, confusion between *read* and *write*, which are both plausible but not equally likely
- all the models struggle in recognizing subtle differences between the plausibility of the situations

Error Analysis

- SDM is the only one picking the right verb for a couple of Traxler pairs:
  - *The hairstylist starts the braid.* (alternative between *combing* and *making*)
  - *The auditor begins the taxes.* (alternative between *auditing* and *doing*)
- all the other models prefer verbs with a more generic and underdetermined meaning (*making, doing*), while SDM chooses the more specific option (*combing, auditing*)

Conclusions

- Logical metonymy interpretation is a hard task, requiring a subtle understanding of event knowledge
- current datasets are challenging even for state-of-the-art models trained on corpora of huge size
- distributional models directly targeting event knowledge (e.g. SDM) can achieve solid performances despite being trained on smaller amounts of data

Conclusions

- Reference: Rambelli G., Chersoni E., Lenci A., Blache P., Huang C.-R. (2020). Comparing probabilistic, distributional and Transformer-based models on Logical Metonymy interpretation. Proceedings of AACL-IJCNLP.

Credits

- Paolo Pedinotti, Giulia Rambelli and Alessandro Lenci (University of Pisa)
- Philippe Blache (Aix-Marseille University)
- Chu-Ren Huang (The Hong Polytechnic University)
- Enrico Santus (MIT / Bayer)

Thank You! Questions?