Do “Future Work” sections have a purpose?
Citation links and entailment for global scientometric questions

Simone Teufel
University of Cambridge and
Tokyo Institute of Technology

August 11, 2017
What is the communicative purpose of these sections?

Other sections in paper are clear:
- Introduction section → Situate yourself and make knowledge claim
- Method section → Claim novelty for method or convince us of correctness

Future work has far less clear communicative purpose.

Additionally, the work is not yet done, so no real contribution can be claimed for it.
Rough instructions

- State what you are currently working on related to this work (and beyond it)
- Make recommendation for possible future research (by others)
- Look into the future, speculate about possible applications
- Mention limitations (and what you are going to do to fix this)
Communicative Purpose

But how does this really help you?

- Show awareness of limitations and potential solutions.
- Convince others that the work must be important because only small problems can be entirely solved.

*We believe that our task formulation of . . . is worthy of additional effort.*

- Warn others off research areas you have already got (invisible) investment already in it.
- Follow social conventions
Risk/Benefit analysis

- **Risk**: If you honestly state your true intentions, somebody can steal your hot ideas before you publish them.
- **Benefit**: Slightly easier to pass peer review.

Rational approach: say as little as you can in Future Work sections.
Alternative Scenario

- Future work sections are written honestly and read with great interest.
- They act as “notice boards” or “hot idea markets” between researchers.
- If this is so, then a lot of work suggested by one set of authors will be taken up by other authors.
Empirical observations

- This talk is not about what the truth is (or how I think “future work” sections should be written).
- It’s about the interesting NLP we work we could be doing finding out what the truth is.
- (I mean supporting a social scientist, historian of science, bibliometrist to find out. )
- Empirically...
Highest variation of all paper sections

- The 10 “Future Work” sections in this workshop...
- ...vary between 25 and 321 words (average of 100.7 words)
- In terms of percentages: 8.6% (max) to 0.5% (min)
- Purely speaking by quantity, people seem to disagree as to the importance of this section.
- But what kinds of things are people writing in their “future work” sections?
FW type 1: Improve results, create resources

- publish the generated abstracts on a publicly available archive (8)
- improve the precision of PDFdigester (2)
- improve the semantic quality of the generated abstracts (8)
- collect a big dataset with the ground truth of recommendation results for training the system (1)
- create a gold standard annotation dataset (subset of representative documents) with curated information on authors, rhetorical categories, citations, etc. (2)
FW type 2: Apply a specific method

- test different weighting functions (e.g., weighting research group size beyond the number of citations) (4)
- measure the similarity between two abstracts using semantic similarity measures (8)
- compute the CPI values as conceptualised by Beel and Gipp in [1] and compare the results by extending the combinations of the metrics like multiplication of CPA and Co-Citation (1)
- extend the similarity metrics using the Word-Net synset hierarchy and distributional similarity and Latent Semantic Analysis (LSA) index (10)
- perform further analysis of other stratagems such as journal run for instance (7)
- use machine learning algorithms using position of the citation as one of the features to improve the weighing process (1)
- experiment with distance measures that reflect the lexico-syntactic structure of language (1)
- introduce more descriptive aspects or domain ontologies (5)
FW type 3: Next Goal/next task

- formalize Choice and Binary type aspects as question-answering tasks (5)
- apply the proposed methodology to support reviewer assignment in the process of paper peer reviewer (4)
- devise and implement a search engine that captures obvious patent abstracts by measuring their similarity to previously generated and published abstracts (8)
- integrate Authors-Topic associations into existing optimization-based strategies (4)
- address community detection and organization finding in the context of project applications (4)
- link problem and solution statements which were found independently during our corpus creation (3)
- (create a) system (that) informs the authors of detected obvious patents (8)
- go beyond log analysis to do user studies in order to compare user feedback with the findings of this study (7)
- investigate the domain specificity of our classifiers and see how well they can generalise to domains other than ACL (e.g. bioinformatics) (3)
FW type 4: Explore and understand

- explore how the concept of “authority” might be applied in this problem (1)
- measure to which degree our classifiers can generalise, i.e., find implicit statements too (3)
- understand the impact of document length on co-citation analysis approaches (1)
- understand the meaning of these differences (6)
- process existing corpora and study the properties of citation contexts on a large scale (9)
- perform user-centric evaluations of our web-based interface in order to better understand the value and possibilities of the rich scientific corpora search and browsing patterns we propose (2)
Uptake of suggestions

- How many suggestions in “Future Work” sections ever get taken up by later papers?
- Answer should be quantified, based on large data set, precise
- Match two things against each other:
  - descriptions of planned research in an earlier set of papers
  - research contributions of later papers
- For instance:
  - *We intend to investigate more sophisticated ways of document representation and of extracting a citation’s context.*
  - *Context Matters: Towards Extracting a Citation’s Context Using Linguistic Features*
  - (from two papers by Daniel Duma and Ewan Klein 2015 and 2016)
Citation Function Detection and Sentiment detection

- Sentiment detection and citation function classification (Nanba and Okumura, 1999); Garzone and Mercer (2000); Nakov et al. (2004); Teufel et al. (2006); Kaplan and Tokunaga (2009); Athar and Teufel (2012); Ding et al. (2014); Catalini et al. (2015); Jha et al. (2017)
This is a proper text understanding task.

My definition of “text understanding”: any process that obtains new knowledge, i.e., something that is true and relevant but that isn’t explicitly stated in the text.

This task involves paraphrasing, citation link processing, citation block determination, and inference (or entailment detection)
Paraphrase detection vs inference

1. *We propose to detect discourse segments by using more sophisticated metrics of coherence.*

2. *Our recogniser finds regions of topically related text . . . (paraphrase)*

3. *We determine gaps between coherent areas of discourse . . . (inference)*

- Understanding how inference works is part of artificial intelligence.
Entailment

A proposition P is said to entail another proposition Q if the truth of Q is a logically necessary consequence of the truth of P (and the falsity of P is a necessary consequence of the falsity of Q).

- Sentence P “That is a dog” entails sentence Q “That is an animal”.
- This means in every situation where I can say P I can also say Q:
  - It can’t possibly be a dog and not an animal.
  - It’s a dog, therefore it’s an animal.
  - If it is not an animal, then it follows that it’s not a dog.
  - ? It’s not an animal, but it’s just possible that it’s a dog.
  - ? It’s a dog, so it might be an animal.
Shared entailment tasks

- Fracas (Cooper et al. 1996)
- Recognising Textual Entailment, RTE (Dagan et al. 2006)
- SNLI (Stanford Natural Language Inference) corpus (Bowman et al. 2015)
FRACAS; 1996

• Designed to cover a wide range of semantic and logical phenomena

<table>
<thead>
<tr>
<th>Positive example</th>
<th>Negative example</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>The really ambitious tenors are Italian.</em></td>
<td><em>No delegate finished the report.</em></td>
</tr>
<tr>
<td>⇒</td>
<td>⊬</td>
</tr>
<tr>
<td><em>There are really ambitious tenors who are Italian.</em></td>
<td><em>Some delegate finished the report on time.</em></td>
</tr>
</tbody>
</table>
Recognising Textual Entailment (RTE); 2005

- Premise taken from naturally occurring text

**The purchase of Houston-based LexCorp by BMI for $2Bn prompted widespread sell-offs by traders as they sought to minimize exposure. LexCorp had been an employee-owned concern since 2008.**

<table>
<thead>
<tr>
<th>Positive example</th>
<th>Negative example</th>
<th>Neutral example</th>
</tr>
</thead>
<tbody>
<tr>
<td>⇒ BMI acquired an American company.</td>
<td>⇓ BMI bought employee-owned LexCorp for $3.4Bn.</td>
<td>⇒ BMI is an employee-owned firm.</td>
</tr>
</tbody>
</table>
SNLI (Stanford NL Inference corpus); 2015

- Situations are grounded in visual scenes/captions
- Crowd-sourced; two separate steps
- Very large (570K pairs)

<table>
<thead>
<tr>
<th>Positive example</th>
<th>Negative example</th>
<th>Neutral example</th>
</tr>
</thead>
<tbody>
<tr>
<td>⇒</td>
<td>⇐</td>
<td>⇒?</td>
</tr>
<tr>
<td><em>Two dogs are running through a field.</em></td>
<td><em>There are animals outdoors.</em></td>
<td><em>The pets are sitting on a couch.</em></td>
</tr>
<tr>
<td><em>There are animals outdoors.</em></td>
<td><em>Some puppies are running to catch a stick.</em></td>
<td></td>
</tr>
</tbody>
</table>

SNLI (Stanford NL Inference corpus); 2015

- Situations are grounded in visual scenes/captions
- Crowd-sourced; two separate steps
- Very large (570K pairs)
Pragmatic-based inference

- Implicatures and presuppositions are another mosaic piece in the puzzle.
- Implicatures are defined as those statements about the world that
  - are assumed to be true by the speaker
  - are transmitted “along with” the literal message
  - without being explicitly stated.
- Presuppositions are a special kind of implicature, one that is lexically triggered.
Presupposition vs Entailment

- Negation of utterance does not cancel its presuppositions:

**Presupposition – no cancellation**

*She has stopped eating meat.*

Presupposition: She used to eat meat.

*She hasn’t stopped eating meat.*

→ Presupposition survives under negation.

- This distinguishes it from entailment.

**Entailment – cancellation**

*The president was assassinated.*

Entailment: The president is dead.

*The president was not assassinated.*

→ Entailment does not survive under negation.

In a sense, we can consider entailments as “part of what is said”.
Presuppositions in Science

1. *Miller et al.* did not manage to verify whether saturation was reached.

2. *Miller et al.* did not verify whether saturation was reached.

- Did Miller even *attempt* verification?
- Yes, for sentence 1.
- No, for sentence 2.
1. *Miller attempted to model how X works.*

2. *Miller modelled how X works.*

- Gricean maxims: always state the strongest relevant and true statement that you possibly can.
- “modelling” (i.e., attempting to model and then succeeding) is stronger than “attempting to model”.
- Therefore: we conclude that Miller’s attempt is criticised as being (at least partially) unsuccessful.
- Some work on automatic classification of presuppositions and implicatures exists (Tremper and Frank, 2013).
- My student Olesya Razuvayevskaya studies how to interpret “let alone” sentences.
Olesya Razuvayevskaya: “let alone”

*The child can’t even sit, let alone walk.*

- A fortiori argumentation:
  - To prove: not A
  - Situation B less likely than A (but comparable)
  - I can prove that not even B
  - \( \rightarrow \) then definitely not A

- How to find this in real language
- How to turn this into a logic formula
- There is an immediate connection to entailment.
- Five main categories of connectedness + questions about negation and “hidden properties”
- Example: *A plan that rethinks a city, let alone shrinks one, is complex, multi-faceted and damn hard to pin down.*
Abstractness classifier

- More abstract “future work” is less likely to be performed (maybe).

- Example pair:
  - *Future improvements to our proposed system include publishing the generated abstracts on a publicly available archive.* (8)
  - *In parallel, we would like to improve the precision of PDFdigester* (2)
  - *It should be extended to try to understand the meaning of these differences* (6)

- Research in metaphor classification contributes classifiers for the abstractness of phrases (“dark hair” vs. “dark humour”; Turney et al. 2011, Neuman et al. 2013)
Global Scientometric questions

- Describe the emergence and development of a scientific field (Bettencourt and Ulwick (2008), Kiss et al. 2010)
- Determine schools of thought (McCain, 1986; Allen, 1997)
- When do scientific revolutions and paradigm shifts (Kuhn 70) occur (De Langhe, 2017)?
- Identifying the scientific areas where the most innovation currently occurs (Chen et al. 2010, Boyack, 2017)
- Detecting the emergence of scientific ideas (Kuhn et al. 2014; McKeown et al, 2016)
Conclusions

• “Future Work” sections may or may not be a meaningful part of papers:
  • True statement of intentions?
  • Or empty socially-induced exercise?

• One exciting task out of a large set of interesting global scientometric questions

• A text understanding task

• Only citation function determination and paraphrasing is unlikely going to be enough

• We also need to go deeper and address “inference” one way or another.

• Start with small objectively evaluable steps, e.g., implicatures, presuppositions or other general principles that don’t require much world knowledge.
• address community detection and organization finding in the context of project applications (4)
• apply the proposed methodology to support reviewer assignment in the process of paper peer reviewer (4)
• create a gold standard annotation dataset (subset of representative documents) with curated information on authors, rhetorical categories, citations, etc. (2)
• (create a) system (that) informs the authors of detected obvious patents (8)
• collect a big dataset with the ground truth of recommendation results for training the system (1)
• compute the CPI values as conceptualised by Beel and Gipp in [1] and compare the results by extending the combinations of the metrics like multiplication of CPA and Co-Citation (1)
• devise and implement a search engine that captures obvious patent abstracts by measuring their similarity to previously generated and published abstracts (8)
• experiment with distance measures that reflect the lexico-syntactic structure of language (1)
• explore how the concept of “authority” might be applied in this problem (1)
• formalize Choice and Binary type aspects as question-answering tasks (5)
• go beyond log analysis to do user studies in order to compare user feedback with the findings of this study (7)
• improve the semantic quality of the generated abstracts (8)
• improve the precision of PDFdigester (2)
• integrate Authors-Topic associations into existing optimization-based strategies (4)
• introduce more descriptive aspects or domain ontologies (5)
• investigate the domain specificity of our classifiers and see how well they can generalise to domains other than ACL (e.g. bioinformatics) (3)
• link problem and solution statements which were found independently during our corpus creation (3)
• measure the similarity between two abstracts using semantic similarity measures (8)
• measure to which degree our classifiers can generalise, i.e., find implicit statements too (3)
• extend the similarity metrics using the Word-Net synset hierarchy and distributional similarity and Latent Semantic Analysis (LSA) index (10)
• perform further analysis of other stratagems such as journal run for instance (7)
• perform user-centric evaluations of our web-based interface in order to better understand the value and possibilities of the rich scientific corpora search and browsing patterns we propose (2)
• process existing corpora and study the properties of citation contexts on a large scale (9)
• publish the generated abstracts on a publicly available archive (8)
• test different weighting functions (e.g., weighting research group size beyond the number of citations) (4)
• use machine learning algorithms using position of the citation as one of the features to improve the weighing process (1)
• understand the impact of document length on co-citation analysis approaches (1)
• understand the meaning of these differences (6)