News Text Segmentation in Human Perception

Elena Yagunova[1], *Lidia Pivovarova*[2], *Svetlana Volskaya*[1]

[1] Saint Petersburg State University, Saint Petersburg, Russia
[2] University of Helsinki, Helsinki, Finland

Data

- Corpus of Russian news devoted to a visit of Arnold Schwarzenegger to Moscow in October 2010 (the cluster)
- 360 documents, 110 thousand tokens

Methodology of the Experimental and Computational Linguistics

- 4 experiments with informants (25+34+20+21 informants)
- 2 computational experiments

Experiments with informants

1) Extract keywords from the text (25 informants)
2) Determine degree of connectivity between sentences in the text (34 informants)
3) Mark syntagmas as understandable and connected "portions" of the text (20 informants)
4) Scale the degree of connectivity between *all* words in the text or between word and punctuation symbol (21 informants)

Computational experiments

1) Extract keywords from the texts (Tf-iDF)

2) Investigate the coherence and segmentation of the corpus documents (open-source Cosegment tool).

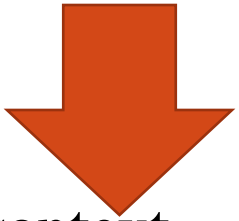
Cosegment produces two outputs: the corpus of texts divided into highly connected segments and the frequency dictionary for these segments.

Results

Keyword extraction

- A set of keywords represents text folding and shows the peculiarities of news texts perception and comprehension by naïve speaker.

- TF-iDF do not extract words which are important for human comprehension because these meaningful words do not distinguish the text in the context of the plot.



- Native speakers use broad context – similar to news stream – to comprehend a particular news message.

Keyword Extraction

Next slide represents information about keywords extracted by informants in comparison with keywords obtained using tf*idf measure (sense vs. cluster characteristics);

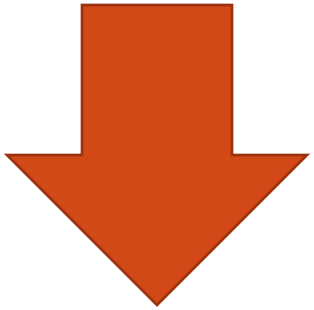
- **the bold font is used to mark the words that appear in the "information portrait" for** the cluster;
- the *italic is used to mark the words that appear in the "information* portrait' in the different part of speech;
- the underline is used to mark the words that appear in the first composition fragment of the text.

Informants	tf*idf
Шварценеггер (Schwarzenegger)	потребовать (to demand)
Медведев (Medvedev)	три-пять (three-five)
технологический (technological)	рубль (rouble)
Сколково (Skolkovo)	Вексельберг (Vekselberg)
долина (Valley)	Russia
Кремниевая (Silicon)	вдвоем (two together)
бум (boom)	*Кремниевый (Silicon)*
ученые (scientists)	установка (aim)
инновационными (by innovation)	целевой (goal)
прорыв (breakthrough)	миллиард (billion)
разработки (products)	объем (volume)
губернатор (governor)	прогноз (forecast)
Арнольд (Arnold)	возможно (perhaps)
российские (Russian)	половина (half)
Дмитрием (by Dmitry)	частный (private)
Калифорния (California)	автомобиль (car)
американских (of American)	бюджетный (budget)
встреча (meeting)	средство (mean)
инновационный (innovative)	течение (stream)
Президентом (by President)	выехать (to leave)
Чайка (Chayka)	общий (common)
я (I)	*президентский (presidential)*
России (Russia)	medvedev@kremlinrussia_e

Results

Discourse segmentation

The majority of keywords (17 of 23) appear in the first narrative component.



Highest weight of the beginning of the text (traditional news text structure).

Discourse segmentation

Governor *of* ***California*** *State* ***Arnold Schwarzenegger*** *considers* ***Russian scientists*** *with the support of the* ***American*** *colleagues to be able to make a* ***technological breakthrough*** *in the* ***innovation*** *center "****Skolkovo****".* ***Schwarzenegger*** *said it during a* ***meeting*** *with* ***Russian President, Dmitry Medvedev****.* ***Schwarzenegger*** *and* ***Medvedev*** *met in the summer of 2010, when the Russian* ***president*** *visited* ***Silicon Valley****. "Then* ***I*** *said to you: "****I*** *will be back." And so* ***I*** *am back," - said the* ***governor*** *of* ***California****. "I was very pleased to hear about your idea to create an equivalent of* ***Silicon Valley*** *in* ***Skolkovo****. Now we will go there, meet with the heads of* ***American investment*** *companies, with their* ***Russian*** *partners. I believe that* ***Russian scientists*** *which are engaged in* ***innovative*** *products, backed by* ***American*** *colleagues can work a miracle, make the* ***technology boom****, "- said* ***Schwarzenegger****. By-turn* ***Medvedev*** *congratulated* ***Schwarzenegger*** *with the fact that California, in fact, is out of the the budget crisis. "****I*** *believe that this is your victory," - noted the* ***president*** *of* ***Russia****. According to him, now in Moscow changes also take place. "We also have a lot of different events. So happens that you have arrived at a time when Moscow has no the mayor," - said* ***Medvedev****. "If you were a citizen of* ***Russia****, you could work with us," - said the head of state, reminding that* ***Schwarzenegger*** *in January 2011 is stepping down as* ***governor*** *of* ***California****. Then* ***Medvedev*** *and* ***Schwarzenegger*** *got into the car "****Chayka****" and left the* ***presidential*** *residence near Moscow in* ***Skolkovo****.*

Results

Syntagmatic segmentation

- Using keywords we estimate supposed weight of each syntagma.
- Sometimes syntagma bounds form a border between topic and focus components.
- Proposition generally coincides with sentence in the news texts, though it could be less than a sentence (e.g., clause).

Syntagmatic Segmentation

- **bold font** is used to highlight the keywords extracted by informants.
- "/" is used to define segment borders
- the segments that do not contain any keywords are crossed out.

Governor *of* ***California*** *State / **Arnold Schwarzenegger** considers / **Russian scientists** / with the support of the **American** colleagues / to be able to make a **technological breakthrough** in the **innovation** center "**Skolkovo**". / **Schwarzenegger** said it during a **meeting** with **Russian President**, **Dmitry Medvedev**. / **Schwarzenegger** and **Medvedev** met in the summer of 2010, / when the Russian **president** visited **Silicon Valley**. / "Then **I** said to you: / "**I** will be back." / And so **I** am back," / - said the **governor** of **California**. "I was very pleased to hear about your idea to create an equivalent of **Silicon Valley** in **Skolkovo**. / ~~Now we will go there~~, / meet with the heads of **American investment** companies, / with their **Russian** partners. / I believe / that **Russian scientists** / which are engaged in **innovative** products, / backed by **American** colleagues can work a miracle, / make the **technology boom**, "- said **Schwarzenegger**. / By-turn / **Medvedev** congratulated **Schwarzenegger** with the fact that California, in fact, is out of the the budget crisis. / "**I** believe that this is your victory," - / noted the **president** of **Russia**. / ~~According to him, now in Moscow changes also take place. / "We also have a lot of different events. / So happens / that you have arrived at a time / when Moscow has no the mayor,"~~ / - said **Medvedev**. / "If you were a citizen of **Russia**, / you could work with us," / - said the head of state, / reminding that **Schwarzenegger** in January 2011 / is stepping down as **governor** of **California**. / Then **Medvedev** and **Schwarzenegger** got into the car "**Chayka**" / and left the **presidential** residence near Moscow in **Skolkovo**. /*

- *Many keywords in syntagma – max weight of syntagma (as type of chunk)*
- *No keywords in syntagma – min weight of syntagma (as type of chunk)*

Results

Text coherence and cohesion

- Computational segmentation in many cases corresponds to the segmentation obtained in psycholinguistic experiment.
- **BUT!** Computational segments are shorter and in some cases non-grammatical.
- This level allows us to describe and classify a text as, for example, a simple event or a sequence of events bound by a cause-effect relation.
- It would be hard to translate the results of these experiments into English because the segmentation is highly depends on micro-syntax structure. Hopefully, the visual clues may give an idea of the potential of this methodology.

Text coherence and cohesion

- **(___)** means the border of connected segment: the connectivity between words within parenthesis is 5 (maximal);
- || means the break, the connectivity is 0 or 1;
- **[ ]** means segmentation obtained in computational experiment with Cosegment program.

(([***Губернатор штата Калифорния***)]([***Арнольд Шварценеггер***] ***считает***)[,*что*] ([***российские ученые***]) ([***при поддержке***]) ([***американских***] [***коллег***) (*смогут совершить*]) *в* ([***инновационном центре***][***"Сколково"***]) ([***технологический прорыв***]). ([***Об этом***]) (***Шварценеггер заявил***) ([***во время***] [***встречи с президентом***] ***России***) ([***Дмитрием Медведевым***]). ***Шварценеггер*** *и* ***Медведев*** [***встречались летом***] ([*2010* ***года***], [***когда***]) ([***российский президент***] ***посещал***) [***Кремниевую долину.***] || ***"Я*** [***тогда (вам сказал***]): (["***Я*** *вернусь*]). [*Вот* ***я и вернулся"***], ([***– сказал ***][***губернатор Калифорнии***]). || "[***Мне*** *было*] (*очень* [*приятно*) (*узнать*] *о*) ([*вашей идее*]) [*создать аналог*]([***Кремниевой долины***]) ([***в Сколково.***]) ***Мы сейчас*** ([***туда поедем***]),([*встретимся с*] [*главами*) ***американских (инвестиционных компаний***]), ***с*** [***их***(***российскими партнерами***]). ([***Я убежден,***] [***что***]) ([***российские ученые***]),([*которые занимаются*) (***инновационными разработками***]), ([***при поддержке***])(***американских*** [***коллег***) (***смогут***) [***совершить чудо***]), [***создать настоящий***(***технологический бум***])[", - ***отметил***] ***Шварценеггер.В*** ([***свою очередь***])(***Медведев*** [***поздравил Шварценеггера***]) ([***с тем***])[, ***что***] ***"Калифорния,*** ([***по*** *сути*]), [*вышла (из*)[*кризиса*) *бюджета*]". ||

Conclusion

- A methodology that combines psycholinguistic and computational experiments applied to the cluster (small homogeneous corpus of news texts) and to one particular document within the corpus. As a result, the document is segmented into structural units of various scales and annotated with operative units crucial for human text comprehension.
- A set of keywords, which is selected during experiment with informants, represents text folding and shows the peculiarities of news texts perception and comprehension by naïve speaker. Quite similar set of keywords can be extracted automatically but this task requires a larger text collection as an input.
- The hierarchy of keywords and their distribution in the text reflect semantic structure of the text: the most important keywords occur in the opening fragment of the text. Comparison of two keywords sets – extracted automatically and selected by informants – characterizes the specificity of the analyzed text in a context of the plot and reflects the peculiarities of perception of the text.
- Keywords (e.g. nominations) in the news texts are often emphasized by syntagma bounds, and sometimes syntagma bounds also form a border between topic and focus components. A proposition generally coincides with sentence in the news texts, though it could be less than a sentence (e.g., clause).
- Propositions join together and form discourse units. This language level allows us to describe and classify a text as, for example, a simple event or a sequence of events bound by a cause-effect relation.

- The results can be further verified by experiments, where informants have to restore the text content using the segments of various length and weight.
- The methodology might be useful
  - i) to compare human perception of various genres and topics
  - ii) to study variety among people in sociolinguistic studies
  - iii) to obtain a "gold standard" for automatic text segmentation;
  - iv) to annotate corpora with cognitive units. However, in two latter cases it would be necessary to implement a lighter version of the method, which would involve fewer informants.

Conclusion

- The Data – the most typical and frequency news text, combine informative and fatic functions

- The methodology can be further define the genre of news text:
  - News texts, combine informative and fatic functions (as in this paper)
  - News texts with informative function

Contacts

Elena Yagunova: iagounova.elena@gmail.com

Lidia Pivovarova: lidia.pivovarova@gmail.com

Svetlana Volskaya: svetlana.volskaya@gmail.com

Acknowledgement: The authors acknowledge Saint-Petersburg State University for a research grant 30.38.305.2014.