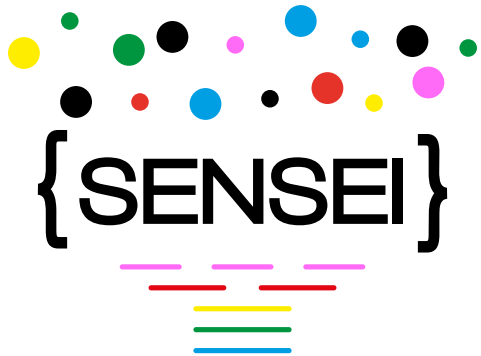
Call-Centre Conversation Summarization pilot task at Multiling'15

Benoit Favre, Evgeny Stepanov, Jérémy Trione, Frédéric Béchet, Giuseppe Riccardi



Aix-Marseille Université, University of Trento

September 3, 2015

Task description

- Create synopses of conversation transcripts
  - Call-centre dialogs
  - Length-limit: 7% of transcript
- Motivation
  - Scenario: quality insurance in a call centre
  - Provide conversation descriptions for browsing a large database of calls
  - Supported by SENSEI FP7 European project
- Objectives
  - Pilot at Multiling'15
  - Speech summarization
  - Conversation modeling
  - Foster research on abstractive summarization
- Multilingual
  - French and Italian transcripts
  - Faithful translations to English

Corpus

- Decoda (French): public transport in Paris, recorded during strikes
  - 1500 dialogs, most of the time 2 speakers
  - Topics: itinerary, schedules, fares, complaints...
- Luna (Italian): company-internal technical help desk
  - 700 dialogs
  - Topics: printer, computer...
- Shared task data
  - 1-5 reference synopses per conversation annotated by SENSEI partners
  - 100 conversations manually translated through SENSEI, the rest with machine translation
  - 500-1000 unannotated
- Test set

Statistic	FR	EN
Conversations	100	100
Turns	7,905	7,909
Words	42,130	41,639
Average length	421.3	416.4
Lexicon size	2,995	2,940
Number of synopses	212	227
Average synopsis length	23.0	26.5

Statistic	IT	EN
Conversations	100	100
Turns	4,723	4,721
Words	34,913	32,502
Average length	349.1	325.0
Lexicon size	3,393	2,451
Number of synopses	500	500
Average synopsis length	17.4	15.4

Examples

- Decoda 20091112-RATP-SCD-0042

Agent: ⟨name⟩ hello
Caller: yes hello
Agent: hello madam
Caller: are buses uh 172 and 186 running?
Agent: unfortunately on the 172 and uh 186, we got the information this morning, there's a notice from the B depot in Vitry so it was known uh yesterday evening and this morning
Caller: uh yes
Agent: so the buses are very disrupted uh this morning huh some uh, uh have gone out and others not, so there are very major disruptions on these two bus lines huh
Caller: whew that's really irritating because what will people who are working do
Agent: unfortunately yeah, it's annoying huh I understand that uh actually
Caller: further there was a notice that was uh
Agent: frankly not uh
Caller: which in fact creates in the private... who risk their post, if they're not going to work because those gentlemen have decided to strike
Agent: it's me I somewhat agree with you
Caller: someone from the RATP who agrees with me
...

- Reference synopses

A1 Are buses 172 and 186 running? No, disrupt because of Vitry depot strike, complaint and compassion
A2 Query of information on the status of buses 172 and 186. Major disruption on these lines due to a strike. Complaint from the caller.

Examples

- Luna 070724-0001

Agent: Okay, could you give me the RWS of your computer?
...
Agent: Okay, if you could stop for a moment, because I still see the lake, ah, alright... okay, show me.
Caller: I'm trying to open my email but it won't open it for me now. The hour glass remains there for a while and then this message appears, saying... impossible ah ...no! Now it's opened it for me and all morning it wouldn't open. No, it's not possible!
Agent: It happens, madam.
Caller: You don't... you wouldn't believe how...
Agent: No, no, I believe it. I believe it, don't worry, it's not... it was for other matters, but it's not the first time...
Caller: Well...
Agent: ...that this happens...
Caller: this has opened it then... so the Internet should also be working... let's see.
Agent: Yes, the Internet should also work.
...

- Reference synopses

A1 The client called because of an error with the mailbox, the agent enters remotely but the problem disappears. Reassures him by saying that it happens.
A2 The client can not open email with Lotus. Telephone support is received. Problem is solved.
A3 The client can not access Lotus web.
A4 the client calls because of problems with opening mail, agent tries to connect but the problem is gone on its own.
A5 The client is not able to access mail. Problem is gone during the phone call.

Submissions

- Constraints
  - 7% of words of original conversation
  - Can submit in any language (EN, FR, IT)
  - Up to 3 submissions
  - Author-provided description of systems available on Multiling website.
- Participants
  - `NTNU:1`: Vector-space model + sentence compression
  - `NTNU:2`: Word representations (CBOW from Mikolov et al.) + sentence compression
  - `NTNU:3`: Sentence representations (paragraph vectors from Le et al.) + sentence compression
  - `LIA-RAG:1`: Graph-based sentence selection with JSD metric, with speech-specific processing
- Baselines
  - `Baseline-MMR`: maximal marginal relevance ($\lambda = .7$)
  - `Baseline-L`: longest turn in conversation
  - `Baseline-LB`: longest turn in 25% first turns of conversation

Evaluation

- Rouge evaluation[1]
  - $\Longrightarrow$ Word n-gram recall between submission and set of reference synopses
    - Version: 1.5.5
    - Crop summary to 7% of conversation words
    - No lemmatization, no stopword removal
    - Jacknifing

- Rouge-2 results

System	EN	FR	IT
NTNU:1	0.023	0.035	0.013
NTNU:2	**0.031**	0.027	0.015
NTNU:3	0.024	0.034	0.012
LIA-RAG:1	-	0.037	-
Basline-MMR	0.029	0.045	0.020
Basline-L	0.023	0.040	0.015
Basline-LB	0.025	**0.046**	**0.027**

[1] Options: `-a -l 10000 -n 4 -x -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0`

Analysis

- Annotator consistency
  - Remove one annotator, then evaluate like system
  - Human: 1-5 for Italian, A-G for French
  - Variance much larger than for systems (~ 0.005)

Annotator	FR	IT
human-1	-	0.121 $\pm$0.023
human-2	-	0.213 $\pm$0.023
human-3	-	0.175 $\pm$0.022
human-4	-	0.073 $\pm$0.014
human-5	-	0.125 $\pm$0.018
human-A	0.194 $\pm$0.029	-
human-B	0.207 $\pm$0.036	-
human-D	0.077 $\pm$0.048	-
human-F	0.057 $\pm$0.039	-
human-G	0.113 $\pm$0.054	-

- Impact of machine translation
  - Split data according to machine translation/manual
  - Automatic translation is more consistent

Annotator	EN-man	EN-auto
NTNU:1	0.018	0.023
NTNU:2	0.019	0.031
NTNU:3	0.015	0.024
Baseline-MMR	0.024	0.033
Baseline-L	0.015	0.030
Baseline-LB	0.023	0.027

Outlook

- Better evaluation
  - SENSEI is piloting extrinsic evaluation with QA supervisers in call centre
  - Are synopses useful for QA supervisers work?
  - Could add community effort: pyramid-like manual evaluation
- True English call centre data
  - Hard to come by. Anyone can contribute a corpus?
  - Might generalize to *anything speech* (meeting recordings, broadcast news, movies), or *anything conversational* (forums, tweets...)
- What's next?
  - Foster more participation, get actual abstractive system participation
  - Will reiterate at next Multiling if enough interest
- Thank you! Questions?