PhD position on Context-aware Machine Translation of User Generated Content.

In the context of a collaborative project between LIMSI’s Spoken Language Processing Group [1] and the INRIA’s AlMaNaCH team[2], we offer a PhD position dedicated to developing data-driven methodologies for context-aware machine translation.

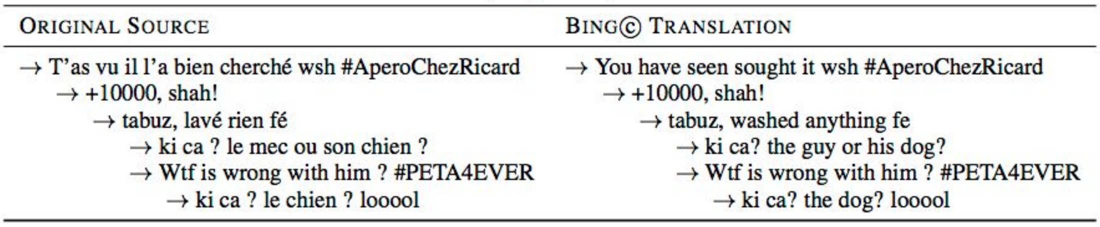
Context

One of the most striking influences of social media on society is how they evolved to impact our perception of events. For instance, during the various Spring Revolutions, FACEBOOK users were in the front line of the information war; more recently, during the November 2015 Paris Attacks, TWITTER was used to gather information about the victims and to offer shelter to those stranded by these attacks. These events generated a steady flow of global textual interactions, crucially highlighting the lack of accurate tools to automatically process and understand these information streams.

Goal

Social media and other forms of online communication have triggered the emergence of new forms of written texts and increase the volume of multilingual user generated content. Translating unlimited streams of highly contextualized and non-canonical texts automatically still remains a scientific challenge. For this, we will have to identify what kind of contextual cues can help translation, in speech situations similar to the one depicted in the figure below[6], and explore new neural architectures for representing them as well as the combination of different levels of information required to deal with the extreme form of ‘noise’ found in UGC.
Note that the project has a strong multilingual orientation (focusing on French, North-African Arabic dialects and English), so the building of robust and language independent model is crucial.

(@rigolboche)



ORIGINAL SOURCE	BING© TRANSLATION
→ T’as vu il l’a bien cherché wsh #AperoChezRicard	→ You have seen sought it wsh #AperoChezRicard
→ +10000, shah!	→ +10000, shah!
→ tabuz, lavé rien fé	→ tabuz, washed anything fe
→ ki ca ? le mec ou son chien ?	→ ki ca? the guy or his dog?
→ Wtf is wrong with him ? #PETA4EVER	→ Wtf is wrong with him ? #PETA4EVER
→ ki ca ? le chien ? looool	→ ki ca? the dog? looool

Typical social media thread initiated by a seed photo and its automatic translation - *Inspired from a real conversation during the last Paris demonstration. Bing was used as it is the official MT engine for Twitter and Facebook.*

The PhD position is funded by the ANR Parsiti project and will be supervised by Djamé Seddah [4] and Guillaume Wisniewski [5].

Profile: MsC in natural language processing or machine learning. An interest in speech processing and/or computer vision would be a plus.
Location: LIMSI - CNRS, Orsay, France
Duration: 36 months PhD
Salary: according to CNRS/INRIA salary grids (including social security, unemployment and retirement benefits)
Deadline for application: open until filled.
Contact: Djamé Seddah (djame.seddah@paris-sorbonne.fr) and Guillaume Wisniewski (guillaume.wisniewski@limsi.fr) with CV + cover letter + reference letter(s)

[1] https://www.limsi.fr/en/research/tlp
[2] https://team.inria.fr/almanach/
[3] https://parsiti.github.io/
[4] http://pauillac.inria.fr/~seddah/
[5] https://perso.limsi.fr/wisniews/
[6] http://pauillac.inria.fr/~seddah/context.pdf