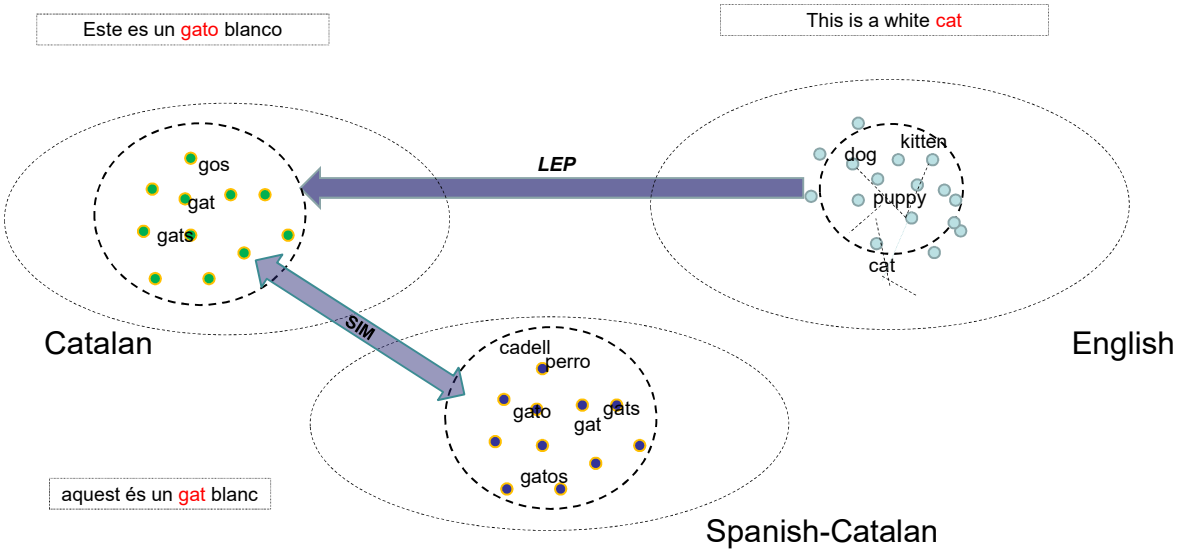
Synthetic Data for Neural Machine Translation of Spoken-Dialects

Hany Hassan, Mostafa Elaraby, Ahmed Tawfik
Microsoft AI & Research

Challenge

- Neural Machine Translation requires large amount of parallel training data.
- Spoken dialects:
  - Not enough parallel data.
  - Widely adopted on social media without standardized written form.
  - Monolingual data mostly with non-standardized written forms.
  - Examples: Arabic spoken dialects, Singaporean-English (Singlish) and etc..

Solution

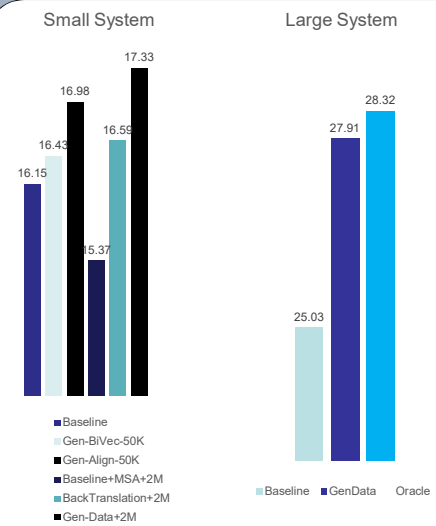


Assume we want to generate Catalan-English from Spanish-English Data

Approach

- Transform data from standard form to written form
- Requires:
  - Word embedding of monolingual data
  - Seed Bi-Lexicon between spoken dialect and English or Spoken and standard dialect.
  - Or use Bilingual embedding BiVec
- Identify local cluster of a given word using kNN queries.
- LEP: Linear Embedding Projection of localized clusters of words to the dialect space.
- SIM: re-rank candidate words by similarity to standard words aligned to a given English word.
- Experimenting on Levantine-Arabic to English using standard Arabic data

Results



Source	النفسية بتختلف كثير مثل لما تكون حاطه بصحن كبير بعدين بتحطه بكاسة
GNMT	Psychological differ many as if it was surrounded by a large dish after two
Our System	The psychological differs a lot like when it's put in a big dish then it is in a cup
Reference	The psychological situation differs a lot when you're put in a huge plate then suddenly you're placed in a glass.
Source	مثلا بيفضلوا أنه أنا أكون مجوزة حدا من ذات البيئة تبعي.
GNMT	For example, they would prefer that I be consumed by the same environment
Our System	For example they prefer that I am married to one of the same environment
Reference	For example they prefer I had married someone from my environment.