Skip to main content

> cs > arXiv:1604.00788

Search... GO

quick links

- Login
- Help Pages
- About

Computer Science > Computation and Language

arXiv:1604.00788 (cs)

[Submitted on 4 Apr 2016 (v1), last revised 23 Jun 2016 (this version, v2)]

Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models

Minh-Thang Luong, Christopher D. Manning

Download PDF

Nearly all previous work on neural machine translation (NMT) has used quite restricted vocabularies, perhaps with a subsequent method to patch in unknown words. This paper presents a novel word-character solution to achieving open vocabulary NMT. We build hybrid systems that translate mostly at the word level and consult the character components for rare words. Our character-level recurrent neural networks compute source word representations and recover unknown target words when needed. The twofold advantage of such a hybrid approach is that it is much faster and easier to train than character-based ones; at the same time, it never produces unknown words as in the case of word-based models. On the WMT'15 English to Czech translation task, this hybrid approach offers an addition boost of +2.1-11.4 BLEU points over models that already handle unknown words. Our best system achieves a new state-of-the-art result with 20.7 BLEU score. We demonstrate that our character models can successfully learn to not only generate well-formed words for Czech, a highly-inflected language with a very complex vocabulary, but also build correct representations for English source words.

Comments: 11pages, 4 figures. ACL 2016 camera-ready version. SOTA WMT'15 English-Czech 20.7 BLEU (+2.1-11.4 points)

Subjects: **Computation and Language (cs.CL)**; Machine Learning (cs.LG)
Cite as: **arXiv:1604.00788 [cs.CL]**
(or **arXiv:1604.00788v2 [cs.CL]** for this version)
https://doi.org/10.48550/arXiv.1604.00788
ⓘ Focus to learn more
arXiv-issued DOI via DataCite

Submission history

From: Minh-Thang Luong [view email]
[v1] Mon, 4 Apr 2016 09:30:54 UTC (53 KB)
[v2] Thu, 23 Jun 2016 00:50:19 UTC (60 KB)
○ Bibliographic Tools

Bibliographic and Citation Tools

□ Bibliographic Explorer Toggle
Bibliographic Explorer *(What is the Explorer?)*
□ Litmaps Toggle
Litmaps *(What is Litmaps?)*
□ scite.ai Toggle
scite Smart Citations *(What are Smart Citations?)*

● Code, Data, Media

Code, Data and Media Associated with this Article

□ Links to Code Toggle
CatalyzeX Code Finder for Papers *(What is CatalyzeX?)*
□ DagsHub Toggle
DagsHub *(What is DagsHub?)*
□ Links to Code Toggle
Papers with Code *(What is Papers with Code?)*
□ ScienceCast Toggle
ScienceCast *(What is ScienceCast?)*
○ Demos

Demos

□ Replicate Toggle
Replicate *(What is Replicate?)*
□ Spaces Toggle
Hugging Face Spaces *(What is Spaces?)*
○ Related Papers

Recommenders and Search Tools

□ Link to Influence Flower
Influence Flower *(What are Influence Flowers?)*

□ Connected Papers Toggle
Connected Papers (*What is Connected Papers?*)
□ Core recommender toggle
CORE Recommender (*What is CORE?*)

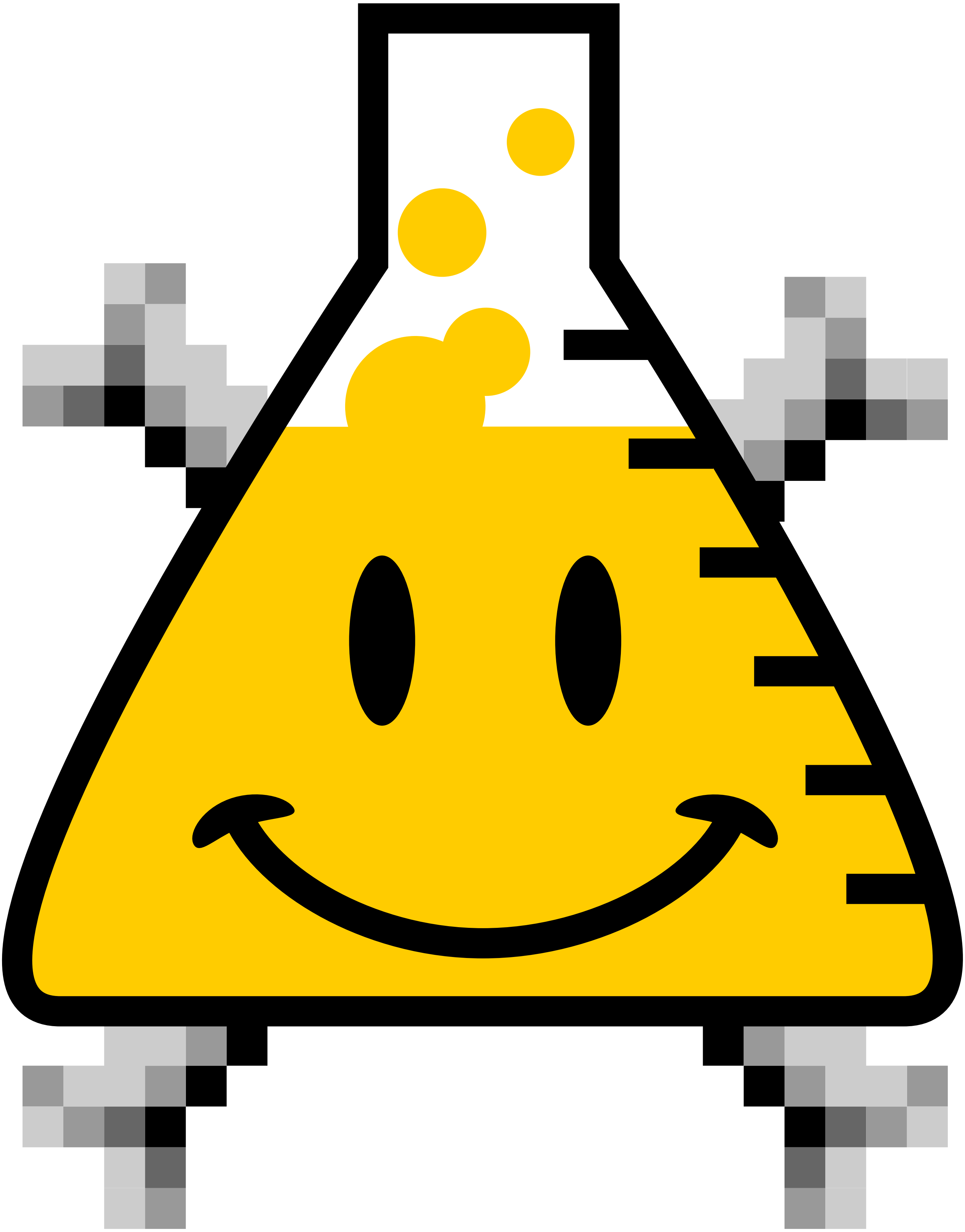
○ About arXivLabs

arXivLabs: experimental projects with community collaborators

arXivLabs is a framework that allows collaborators to develop and share new arXiv features directly on our website.

Both individuals and organizations that work with arXivLabs have embraced and accepted our values of openness, community, excellence, and user data privacy. arXiv is committed to these values and only works with partners that adhere to them.

Have an idea for a project that will add value for arXiv's community? **Learn more about arXivLabs**.



Which authors of this paper are endorsers? | Disable MathJax (What is MathJax?)