Review of thesis proposal

Reviewer: David Mareček, ÚFAL MFF UK

Date: September 17, 2021

Thesis Title: Multi-Lingual Machine Translation

Candidate: Dominik Macháček

Supervisor: Ondřej Bojar, ÚFAL MFF UK

Obsah práce

Cílem práce Dominika Macháčka je prozkoumat možnosti strojového překladu mluvené řeči, kde je k dispozici více zdrojů, tj. například originální řeč v jednom jazyce a její simultánní překlad do jiného jazyka. Po úvodní kapitole následuje přehled metod používaných ve vícejazyčném překladu. Následuje kapitola diskutující problémy simultánního překladu mluvené řeči, jeho metody a evaluaci. Čtvrtá kapitola popisuje data a nástroje, které budou v experimentech využívány. Následuje výčet problémů které v automatickém simultánním tlumočení nastávají a možnosti jejich řešení. V šesté kapitole pak Dominik navrhuje jedno z možných řešení vícezdrojového simultánního překladu, které hodlá implementovat. V sedmé kapitole pak rozděluje svou dizertační práci do tří na sebe závisejících úkolů: a) data, b) baseliny, c) simultánní překlad.

Komentáře k práci a dotazy

1) Plánovaný experiment popsaný v kapitole 6 není popsán úplně do detailů (což je v pořádku vzhledem k formátu tezí). Chápu, že alignment mezi originálním zdrojem a interpretovaným zdrojem se bude učit neřízeně, tj. bude se předpokládat, že když překladové okno bude dost velké, tak si to tam Transformer někde najde sám. K tomu mám několik dotazů:
    a) Pokud se budou používat pro trénování klasická textová data, předpokládám, že je bude třeba upravit tak, aby jeden zdrojový jazyk byl také posunutý vůči tomu druhému. Plánujete něco takového?
    b) Z obrázku 1 to vypadá, že se z obou zdrojů bere vždy posledních X slov. Neuvažovali jste o tom, že by se ten originální zdroj posunul a nebralo se z něj několik posledních slov? Tím by se zvýšil překryv mezi těmi dvěma zdroji ale naopak by se prodloužila latence.

2) Data: Jak velké množství trénovacích dat je možné dostat z EU 2008-2011 pomocí ASR? Předpokládám, že kvalita sice nebude moc dobrá, ale na druhou stranu to bude více simulovat zadaný úkol. Máte představu, zda bude lepší použít ASR spíše jako preprocessing k překladu, nebo ho použít pro výrobu terénovacích dat? Daly by se tyhle dva přístupy nějak zkombinovat?
3) Co se týká navržených milníků, jaké jsou rozdíly mezi offline a simultáním překladem? Budete v tom offline jinak řešit alignment, nebo pouze při evaluaci nebudete řešit latenci?

Závěr

Dominik prokázal, že se dobře orientuje v zadaném tématu. Má přehled o existujících datech, nástrojích i souvisejících pracech. Experimenty, které si naplánoval, jsou reálně proveditelné. Spolu s věcmi, které udělal do teď, bude mít dostatečné množství materiálu k sepsání dizertační práce. Jediné riziko vidím v tom, že jeho vícezdrojový překlad využívající simultánní tlumočení nedokáže porazit vylepšený jednozdrojový baseline. Teze jednoznačně doporučuji k obhájení.