HimL at QT21 Workshop

Barry Haddow

Valencia, March 4th, 2017

Overview

1. Project Overview
2. System releases: from PBMT to NMT
3. Lessons for Researchers
4. Producing Unseen Morphological Variants in SMT
   - Matthias Huck

HimL: Health in my Language

Type:	Horizon 2020 Innovation Action
Duration:	February 2015 to January 2018
Website:	`www.himl.eu`

THE UNIVERSITY of EDINBURGH

LMU, Munich

Charles University, Prague

LINGEA

NHS 24

Cochrane

Health in my Language

Global Objectives

- Increase the accuracy of machine translation, making it more reliable and more widely useful.
- Increase the availability and reliability of local public health information to recent immigrants.
- Increase access to the latest best practices in health care information to people whose first language is not currently well supported.
- Decrease the cost to public services of maintaining large amounts of multi-lingual content.

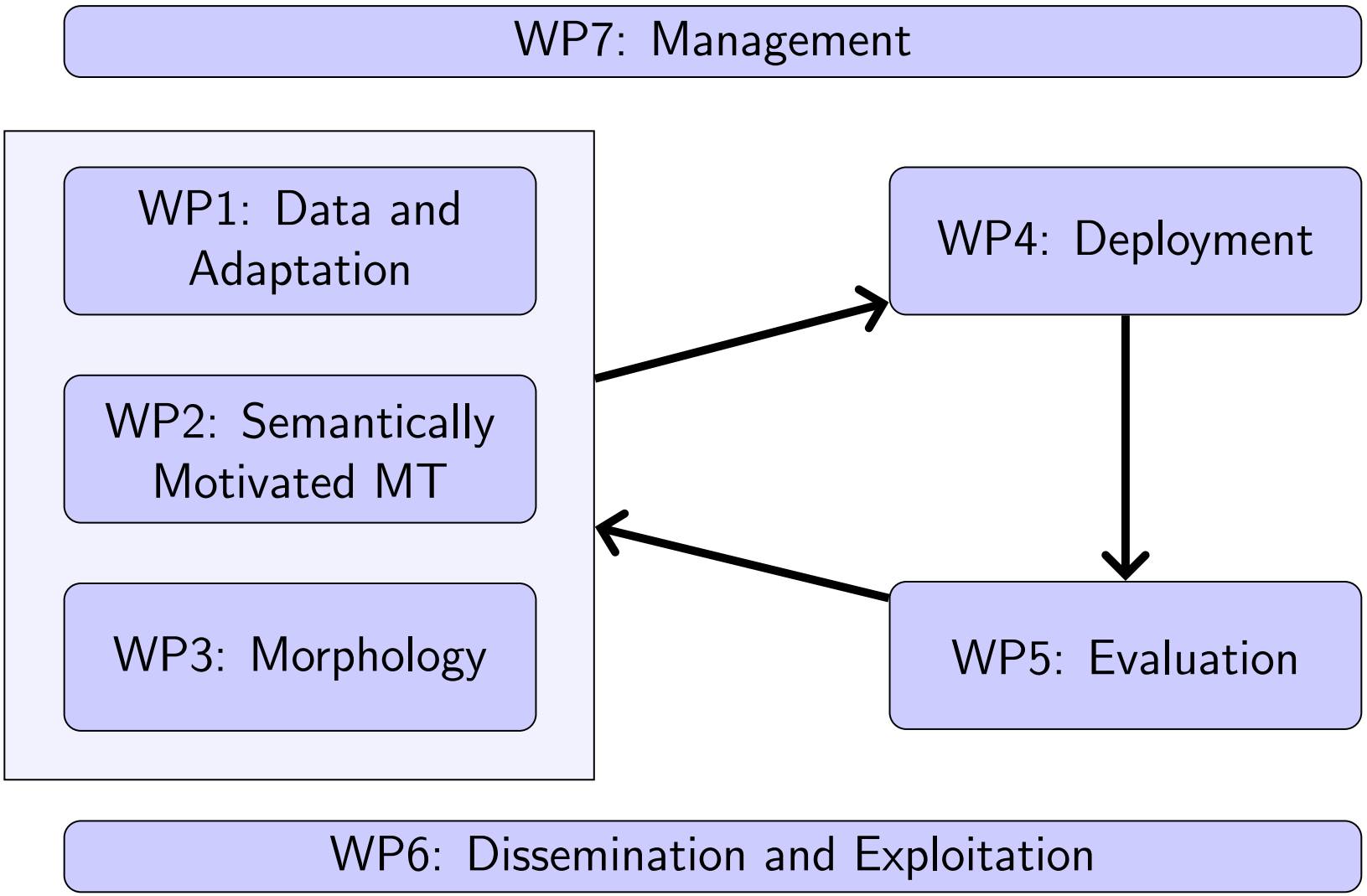
[HimL Proposal, 2014]

Implementational Objectives

- Build accurate MT systems, for public health
- Deploy the systems
- Integrate MT into CMS
- Manage user expectations
- Measure impact

[Summarised from, HimL Proposal, 2014]

Work Packages



Health in my Language

Languages

English

Czech
German
Polish
Romanian

System Release Plan

Year	Languages	Adaptation	Morphology	Semantic	Evaluation
1	de, cs	●	○	○	Human
	ro, pl	○	○	○	Automatic
2	de, cs	●	●	◐	User Acceptance
	ro, pl	●	○	◐	User Acceptance
3	de, cs	●	●	●	Field
	ro, pl	●	●	●	Field

Release Cycle

Month	Phase
July	System release
August-September	Deployment
October-November	Evaluation
December-January	Analysis and Reporting

Y2 System Release

- Baseline
  - Phrase-based (Moses), OSM, hier-reordering
- Data
  - Union of WMT, Khresmoi, OPUS (50-80M sentences)
  - Small Cochrane sets (en-de $\approx$ 10k; en-pl $\approx$ 1k)
  - Project specific dev and test sets
- Adaptation
  - Compared standard Moses techniques
  - LM interpolation, TM interpolation / provenance features
- Morphology
  - en-de uses two-step approach
  - en-cs based on Chimera (including depfix)
- Semantics
  - "Core fidelity" filtering of phrase tables

Health in my Language HimL

Y1/Y2 System Releases: Automatic Scoring

	en-cs		en-de		en-pl		en-ro	
	Coch	NHS	Coch	NHS	Coch	NHS	Coch	NHS
Y1	23.7	20.4	33.8	**30.1**	15.7	23.6	27.8	25.9
Y2	**25.6**	**21.3**	**35.6**	26.2	**17.0**	**25.4**	**37.1**	**32.6**

- en-ro affected by diacritic processing
- en-de Y2 (NHS 24) under investigation

First NMT Systems

- Based on Nematus/DL4MT setup used in WMT16
- Use all parallel data – train to convergence

First NMT Systems

- Based on Nematus/DL4MT setup used in WMT16
- Use all parallel data – train to convergence

	en-cs		en-de		en-pl		en-ro	
	Coch	NHS	Coch	NHS	Coch	NHS	Coch	NHS
Best Y1/Y2	25.6	21.3	35.6	30.1	**17.0**	**25.4**	**37.1**	**32.6**
NMT Base	**30.2**	**23.1**	**37.6**	**31.6**	15.5	19.5	31.5	28.6

Adaptation of NMT Systems

Main Idea

Continued Training (aka Fine-Tuning)

What data to use?

Cochrane – Too small

EMEA – Small, not really in-domain

Synthetic – No target in-domain

Solution

Create synthetic data from common-crawl selection

Adaptation with Synthetic Data

Process

1. Crawl partner website (NHS 24, Cochrane)
2. Translate to target languages
3. Use Moore-Lewis to select from common-crawl
4. Back-translate to create synthetic
5. Continue training with 50-50 synthetic/parallel

	en-cs		en-de		en-pl		en-ro	
	Coch	NHS	Coch	NHS	Coch	NHS	Coch	NHS
NMT Base	30.2	23.1	37.6	31.6	15.5	19.5	31.5	28.6
Select (Coch)	**33.4**	25.6	38.5	31.7	**19.1**	**24.9**	**34.4**	29.0
Select (NHS)	33.2	**26.7**	**39.2**	**32.9**	18.9	24.2	34.1	**29.7**

Health in my Language HimL

Human Evaluation

	en-cs	en-de	en-pl	en-ro
NMT	**1.398**	**1.717**	**0.712**	-0.174
Y2	-0.329	-1.440	-0.579	**0.284**
Y1	-1.238	-0.822	-0.626	-1.586

- Pairwise ranking, scored as in WMT
- Suspected diacritic issues with en-ro data

Health in my Language HimL

Towards Y3 Systems

- Technology
  - Aim to deploy NMT for all language pairs
- Data
  - Released medically-oriented *UFALCorpus*
  - Includes re-crawl of EMEA
  - Investigating effect of noise on en-ro
- Adaptation
  - Incorporate context and domain indicators into NMT
  - Replace back-translation with auto-encoder
- Morphology
  - Application of corrective and separation approaches
- Semantics
  - Potentially bigger problem with NMT?
  - Investigating SRL, round-trip, ...

Health in my Language HimL

Feedback to Researchers – What do Users Need?

- Accurate and appropriate translation, nothing added or removed.
- Grammatically correct translation.
- Correct register, tone, dialect.
- Care with cultural references, which reader may not understand.
- Correct usage of technical terms / jargon.
- Preservation of formatting information (esp. with NMT)
- Assistance with authoring text in “plain English”