Master thesis

The effects of word segmentation quality on word alignments

Student: Ane Berasategi

Supervisor: Masoud Jalili Sabet, Hinrich Schütze

Submitted: August 2020

Motivation

- Create a **pipeline** of BPE + word alignments + mapping from subword alignments to word alignments
- Improvement of the algorithm to learn BPE units
- Improvement of BPE baseline
- Implementation of **chaos mode**: segmentations without word boundaries

Plan

Part 1: BPE pipeline

Part 2: Improvements

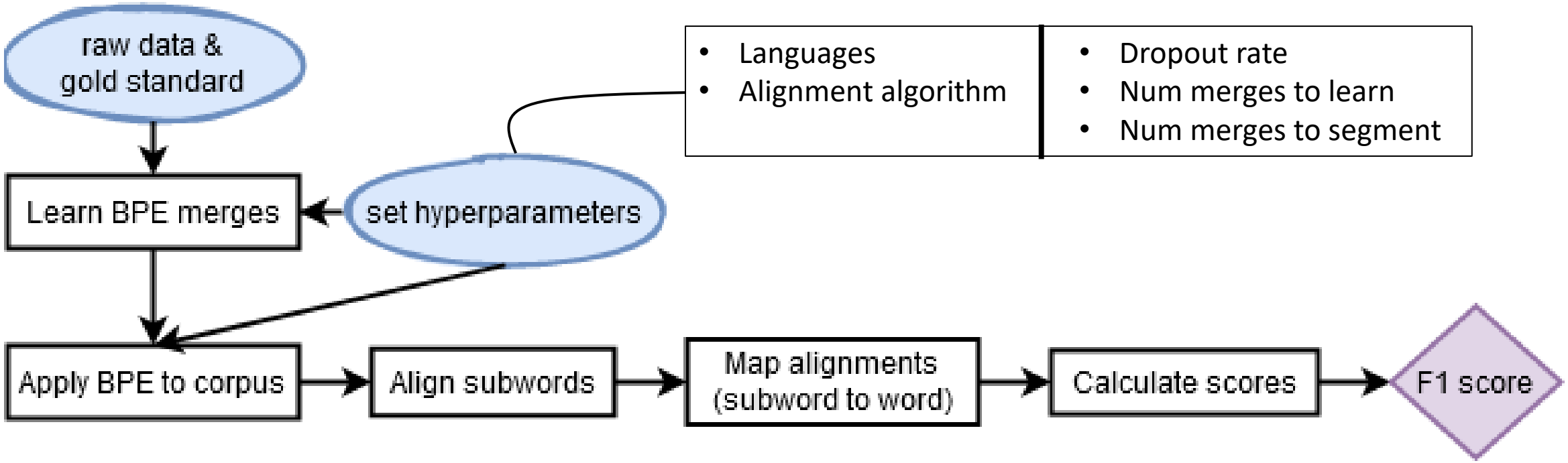
- learn-BPE algorithm
- Chaos mode

Part 3: Results

Part 4: Conclusion

1. BPE pipeline

English sentence after 4000 merges:
_they _will _certain ly _en h ance _the _feeling _of _the _right _of _mo vement



Plan

Part 1: BPE pipeline

Part 2: Improvements

- learn-BPE algorithm
- Chaos mode

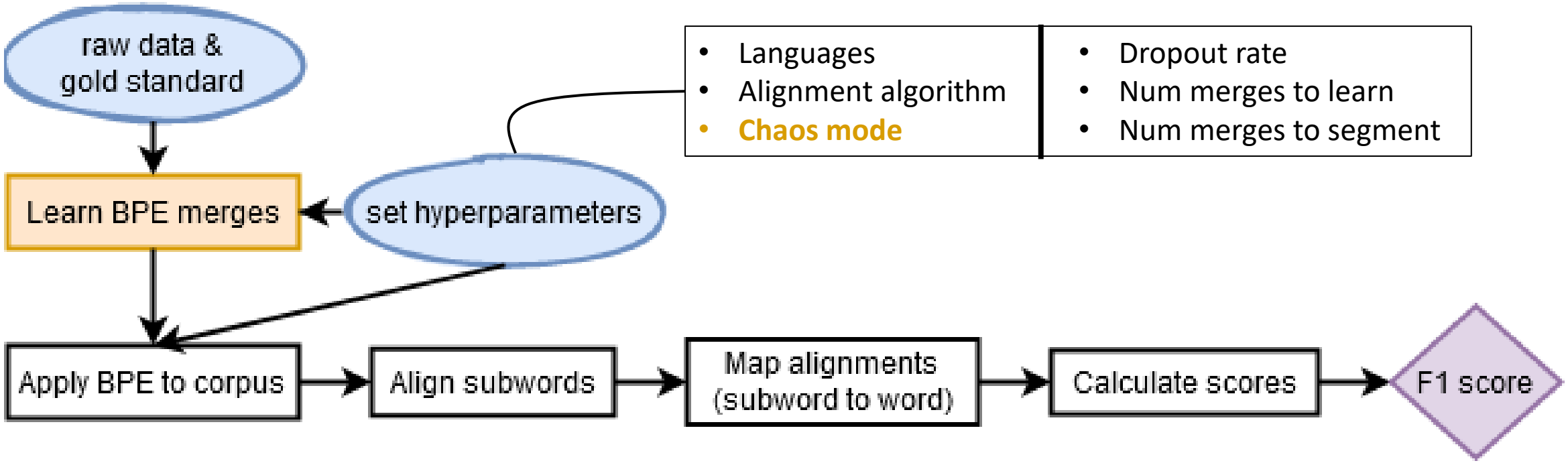
Part 3: Results

Part 4: Conclusion

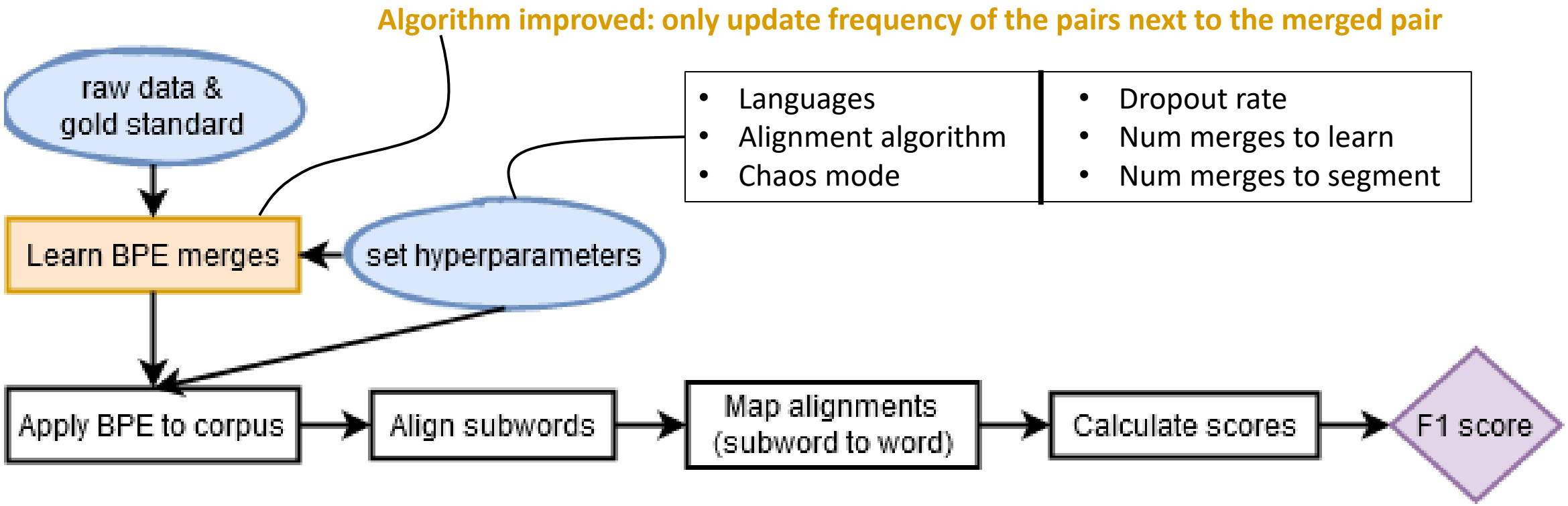
2. Improvements

Chaos mode: segmentations without word boundaries

English sentence after 500 merges (chaos mode):
_they_will _cert ain ly _en h ance _the _feel ing_of_the _ri ght_of _mo vement



2. Improvements



Plan

Part 1: BPE pipeline

Part 2: Improvements

- learn-BPE algorithm
- Chaos mode

Part 3: Results

Part 4: Conclusion

3. Results

- **Space mode**: segmentations with word boundary
- **Chaos mode**: segmentations without word boundary

Best hyperparameters

Dropout rate		Num units to segment		Alignment threshold	
Space	Chaos	Space	Chaos	Space	Chaos
10%	20%	2000	200-500	70%	50%

- Improvement of learn-BPE algorithm: speedup by 2.5x, and it takes 3.5x more time for chaos mode

3. Results

- **Space mode**: segmentations with word boundary
- **Chaos mode**: segmentations without word boundary

Best hyperparameters

Dropout rate		Num units to segment		Alignment threshold	
Space	Chaos	Space	Chaos	Space	Chaos
10%	20%	2000	200-500	70%	50%

- Improvement of learn-BPE algorithm: speedup by 2.5x, and it takes 3.5x more time for chaos mode

Baseline BPE results

Eng – Deu (10k sentences)		Eng – Ron (50k sentences)		Eng – Hin (3k sentences)	
Space	Chaos	Space	Chaos	Space	Chaos
F1: 0.609	F1: 0.477	F1: 0.55	F1: 0.538	F1: 0.381	F1: 0.288

Improvements of BPE-dropout vs. BPE

Eng – Deu		Eng – Ron		Eng – Hin	
Space	Chaos	Space	Chaos	Space	Chaos
F1: 0.635	F1: 0.559	F1: 0.564	F1: 0.58	F1: 0.396	F1: 0.32

5. Conclusion

- BPE + word alignment pipeline
- Proved the improvement of BPE-dropout over BPE in 3 language pairs
- Improvement of learn-BPE algorithm
- Implementation of chaos mode in the pipeline
- Chaos mode: improvement of BPE-dropout over BPE in all languages

Master thesis

The effects of word segmentation quality on word alignments

Student: Ane Berasategi

Supervisor: Masoud Jalili Sabet, Hinrich Schütze

August 2020