Learning multilingual topics through aspect extraction from monolingual entity-centered, opinionated texts

Johannes Huber and Myra Spiliopoulou

IWCLUL 2019

7th of January 2019

Outline

1.

Task and Motivation

What were we trying to do and why?

I think that Tartu University's cafeteria serves great salads.

I think that Tartu University's cafeteria serves great salads.

Opinion Expression

Aspect Term

I will never forget the amazing meal, service, and ambiance at this restaurant.

This place has a cute interior decor and affordable city prices.

The pad seew chicken was delicious, however the pad thai was far too oily.

I will never forget the amazing meal, service, and ambiance at this restaurant.

Paistetut kasvikset olivat lähes raakoja, hinta aika korkea.

This place has a cute interior decor and affordable city prices.

Paras thai-paikka Turussa, maistuvaa ja tuoretta ruokaa, ystävällinen palvelu.

The pad seew chicken was delicious, however the pad thai was far too oily.

meal
pad thai
pad seew chicken
paistetut kasvikset
ruokaa

interior decor
ambiance

Palvelu
service

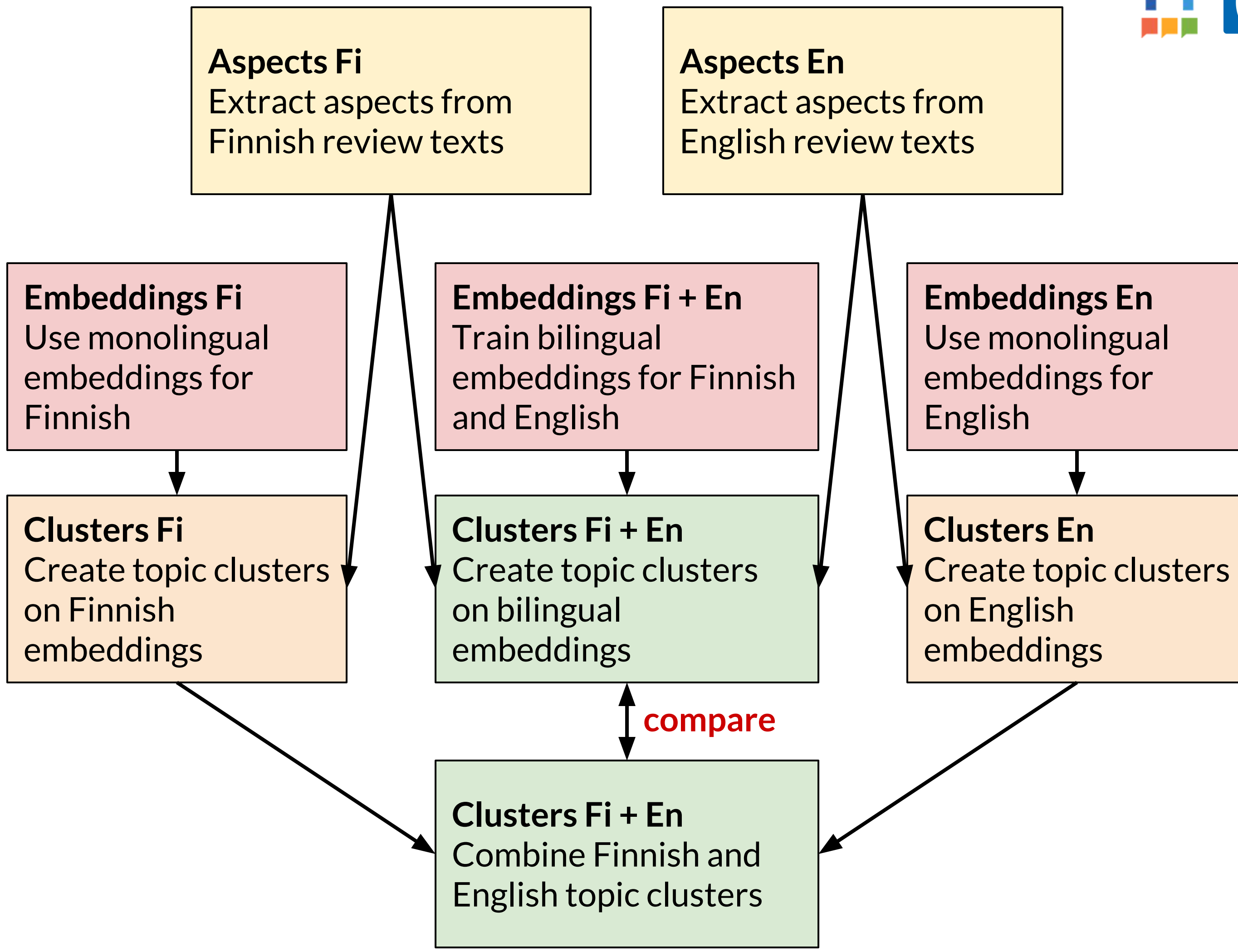
hinta
city prices

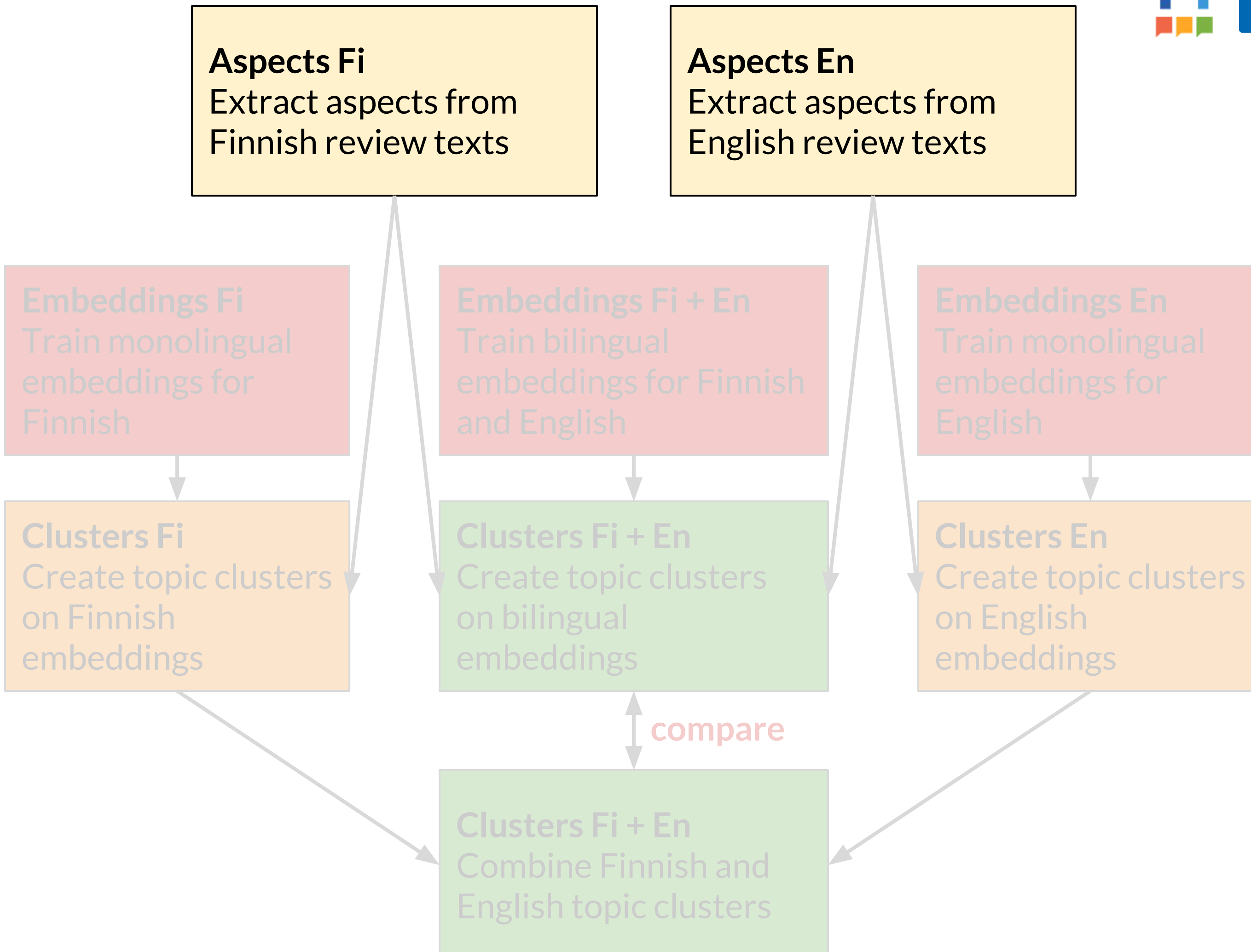
thai-paikka

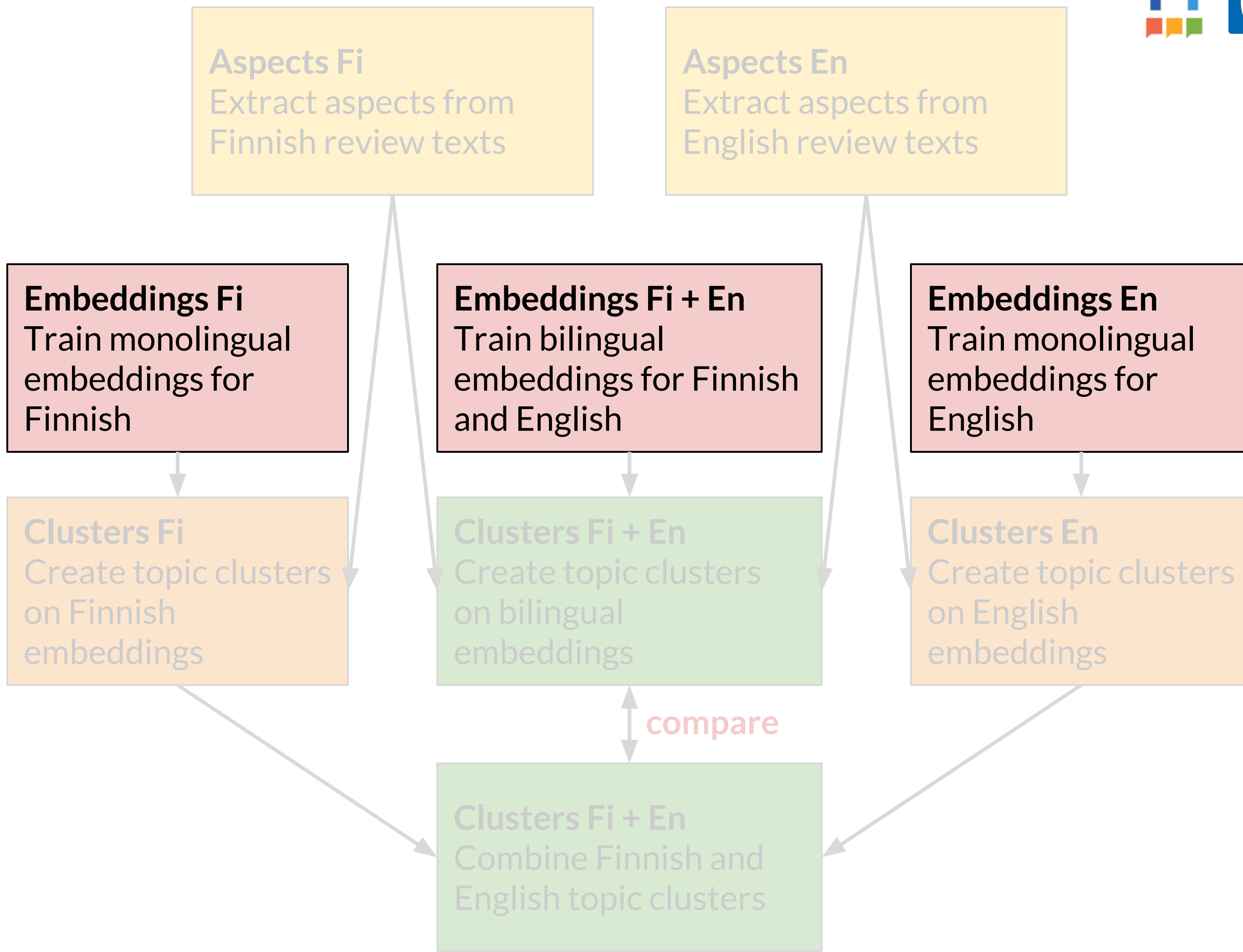
2.

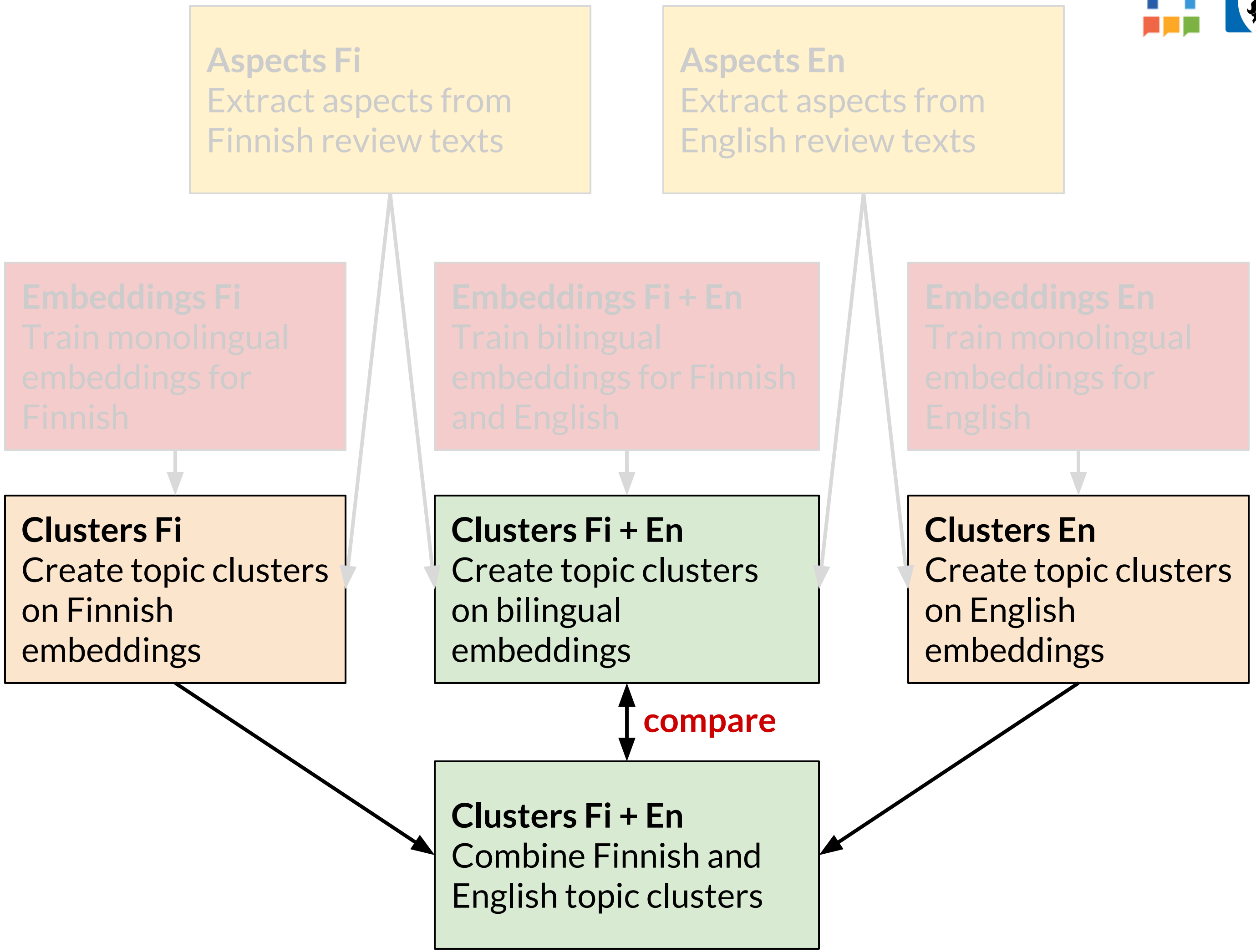
Framework Architecture

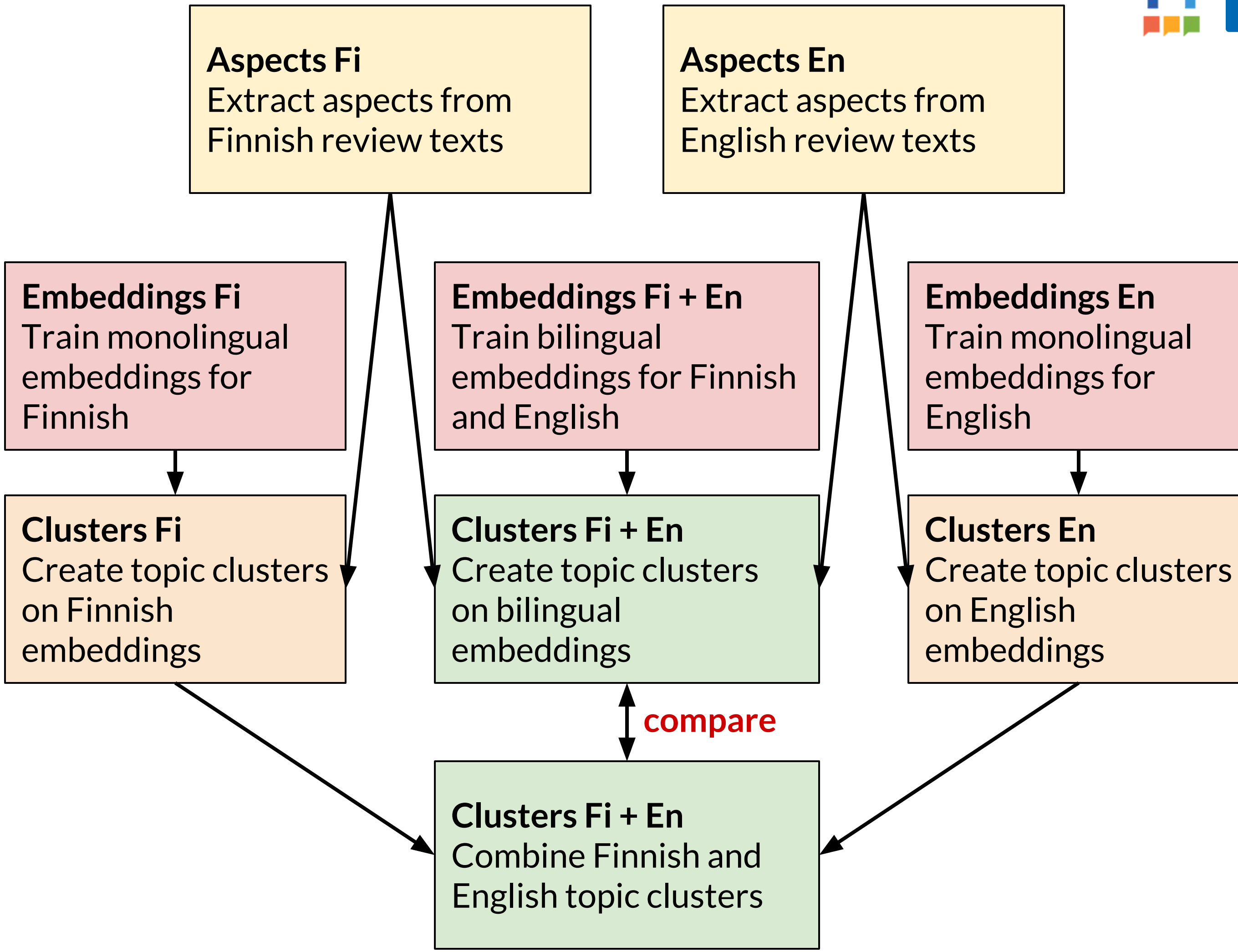
How does the system work?

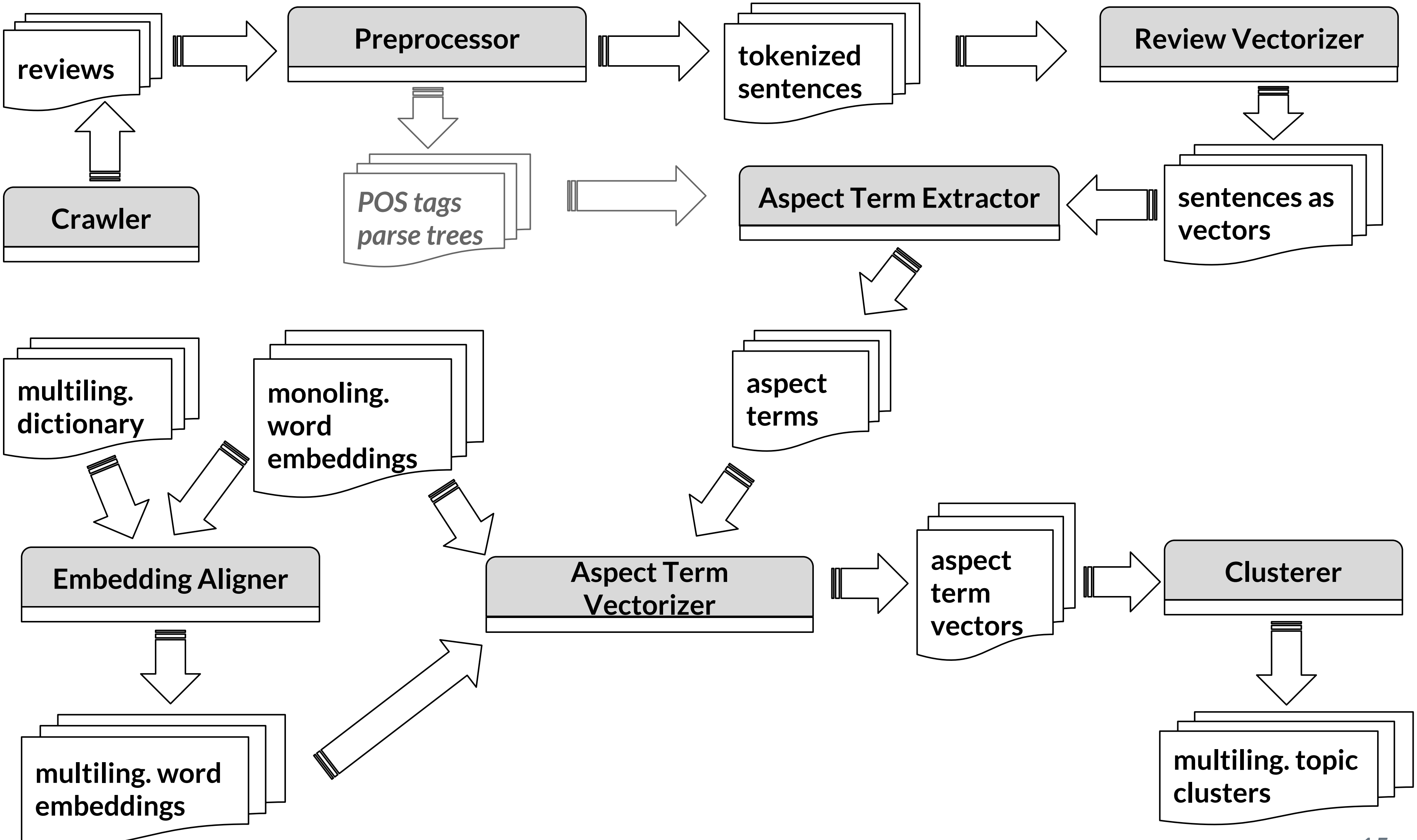

Aspects Fi
Extract aspects from Finnish review texts
Aspects En
Extract aspects from English review texts
Embeddings Fi
Use monolingual embeddings for Finnish
Embeddings Fi + En
Train bilingual embeddings for Finnish and English
Embeddings En
Use monolingual embeddings for English
Clusters Fi
Create topic clusters on Finnish embeddings
Clusters Fi + En
Create topic clusters on bilingual embeddings
Clusters En
Create topic clusters on English embeddings
compare
Clusters Fi + En
Combine Finnish and English topic clusters


Aspects Fi
Extract aspects from Finnish review texts
Aspects En
Extract aspects from English review texts
Embeddings Fi
Train monolingual embeddings for Finnish
Embeddings Fi + En
Train bilingual embeddings for Finnish and English
Embeddings En
Train monolingual embeddings for English
Clusters Fi
Create topic clusters on Finnish embeddings
Clusters Fi + En
Create topic clusters on bilingual embeddings
Clusters En
Create topic clusters on English embeddings
compare
Clusters Fi + En
Combine Finnish and English topic clusters


Aspects Fi
Extract aspects from Finnish review texts
Aspects En
Extract aspects from English review texts
Embeddings Fi
Train monolingual embeddings for Finnish
Embeddings Fi + En
Train bilingual embeddings for Finnish and English
Embeddings En
Train monolingual embeddings for English
Clusters Fi
Create topic clusters on Finnish embeddings
Clusters Fi + En
Create topic clusters on bilingual embeddings
Clusters En
Create topic clusters on English embeddings
compare
Clusters Fi + En
Combine Finnish and English topic clusters

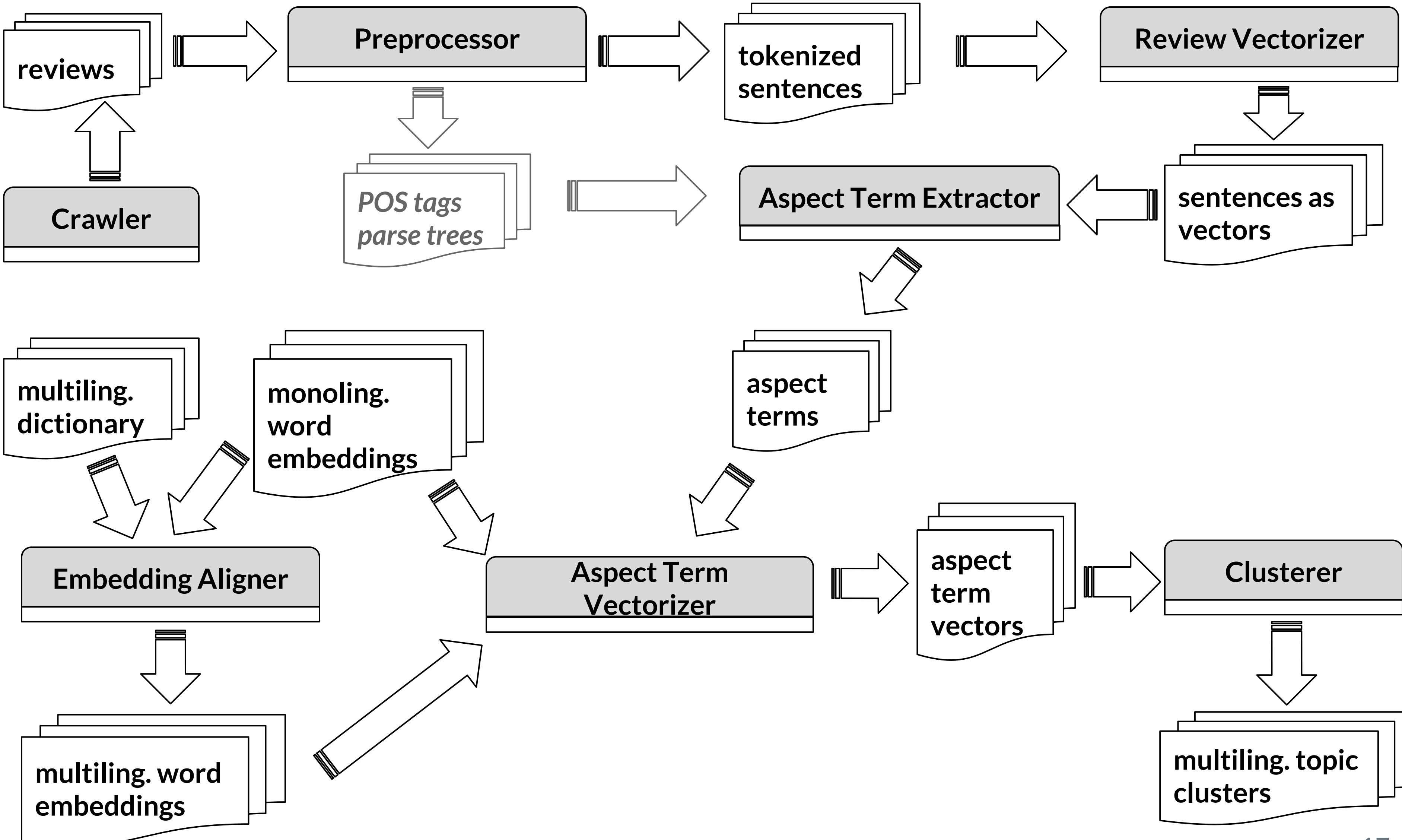

Aspects Fi
Extract aspects from Finnish review texts
Aspects En
Extract aspects from English review texts
Embeddings Fi
Train monolingual embeddings for Finnish
Embeddings Fi + En
Train bilingual embeddings for Finnish and English
Embeddings En
Train monolingual embeddings for English
Clusters Fi
Create topic clusters on Finnish embeddings
Clusters Fi + En
Create topic clusters on bilingual embeddings
Clusters En
Create topic clusters on English embeddings
compare
Clusters Fi + En
Combine Finnish and English topic clusters


Aspects Fi
Extract aspects from Finnish review texts
Aspects En
Extract aspects from English review texts
Embeddings Fi
Train monolingual embeddings for Finnish
Embeddings Fi + En
Train bilingual embeddings for Finnish and English
Embeddings En
Train monolingual embeddings for English
Clusters Fi
Create topic clusters on Finnish embeddings
Clusters Fi + En
Create topic clusters on bilingual embeddings
Clusters En
Create topic clusters on English embeddings
compare
Clusters Fi + En
Combine Finnish and English topic clusters


reviews
Preprocessor
tokenized sentences
Review Vectorizer
Crawler
POS tags parse trees
Aspect Term Extractor
sentences as vectors
multiling. dictionary
monoling. word embeddings
aspect terms
Embedding Aligner
Aspect Term Vectorizer
aspect term vectors
Clusterer
multiling. word embeddings
multiling. topic clusters

Preprocessor

- Detect review language
- Split reviews into sentences
- Tokenize sentences
- For some aspect extraction methods:
    - Lemmatize words
    - Remove stopwords
    - Detect parts-of-speech
    - Get dependency/ constituency parse trees


reviews
Crawler
Preprocessor
POS tags
parse trees
tokenized sentences
Review Vectorizer
sentences as vectors
Aspect Term Extractor
aspect terms
multiling. dictionary
monoling. word embeddings
Embedding Aligner
multiling. word embeddings
Aspect Term Vectorizer
aspect term vectors
Clusterer
multiling. topic clusters

Aspect Term Extraction

- Sequence labelling task

The	*pad*	*seew*	*chicken*	*was*	*delicious.*
O	B	I	I	O	O

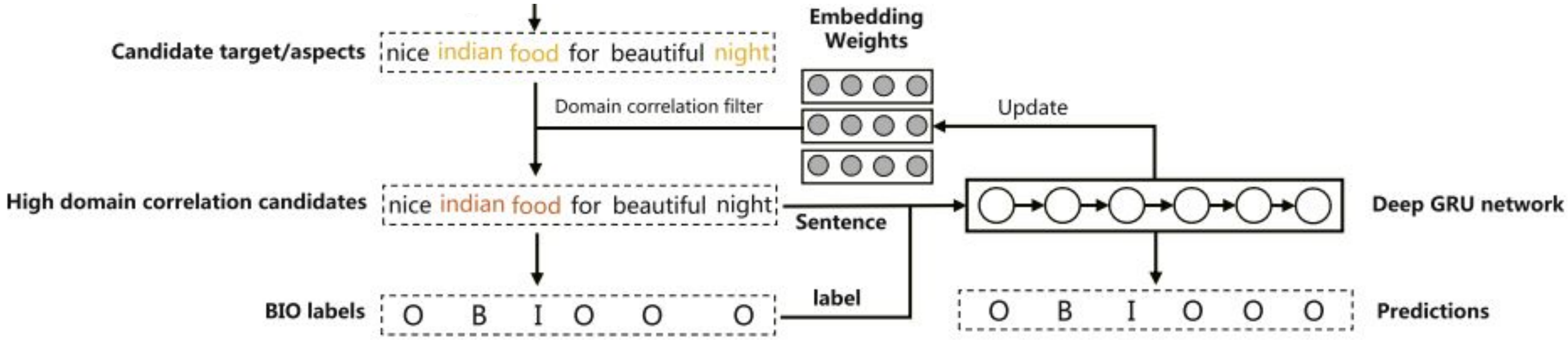
- Compare and optimize three methods
  - Supervised learning
  - Hybrid rule-based/learning
  - Attention-based, unsupervised learning
- Evaluate against gold corpus, calculate F1-Value, choose best performing method for next steps

Supervised Learning

- "Double Embeddings and CNN-based Sequence Labeling for Aspect Extraction" (Xu et al., 2018)
- Double Embeddings:
  - General embeddings trained on huge corpus
  - Domain-specific embeddings
- Train Convolutional Neural Network on sentences represented by double embeddings
- Used without changes for Finnish

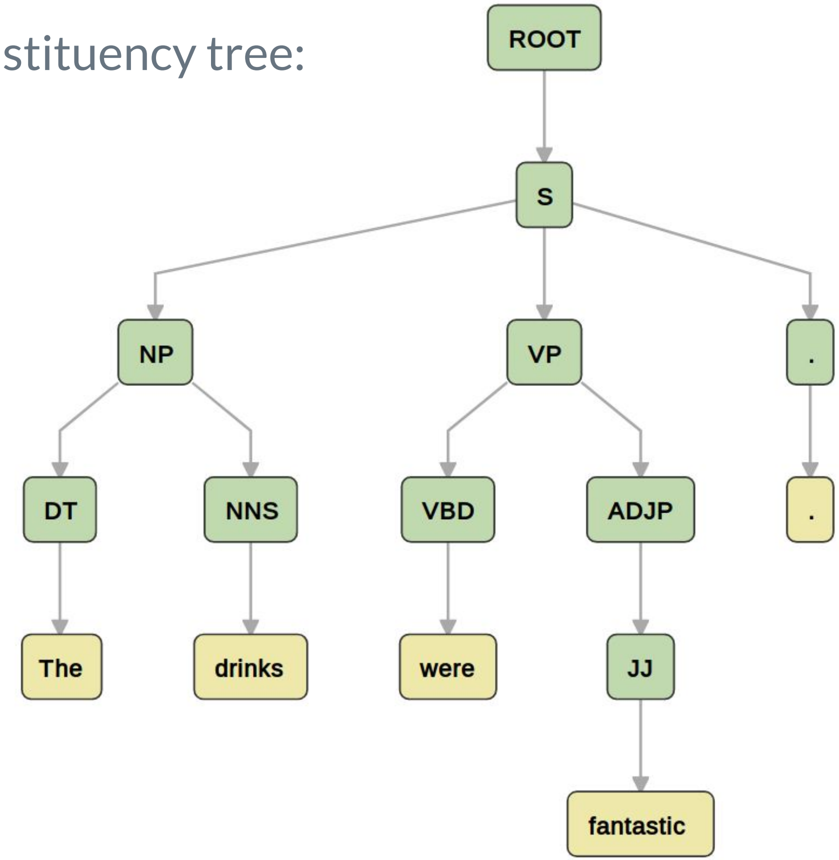
Rule-based hybrid

- “A hybrid unsupervised method for aspect term and opinion target extraction” (Wu et al., 2018)
- Extract candidates via rules from parse tree (=baseline)



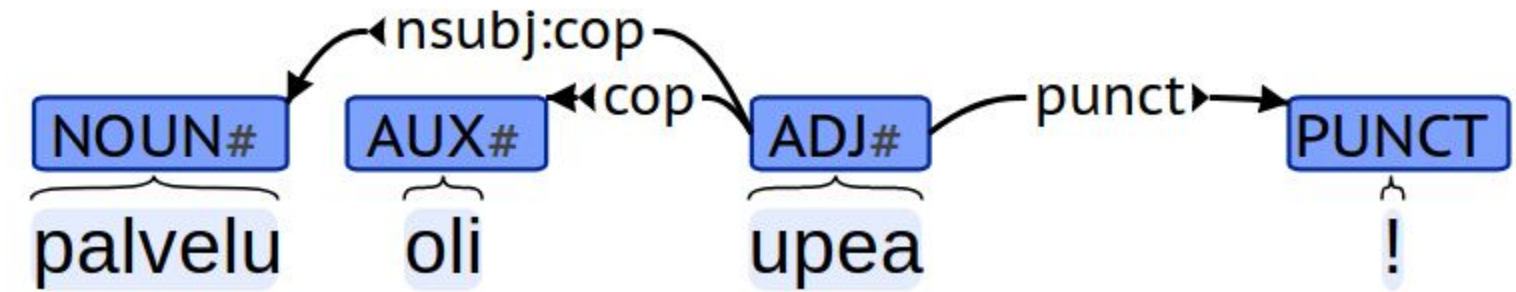
Rule-based hybrid

English constituency tree:



Rule-based hybrid

Finnish dependency tree:

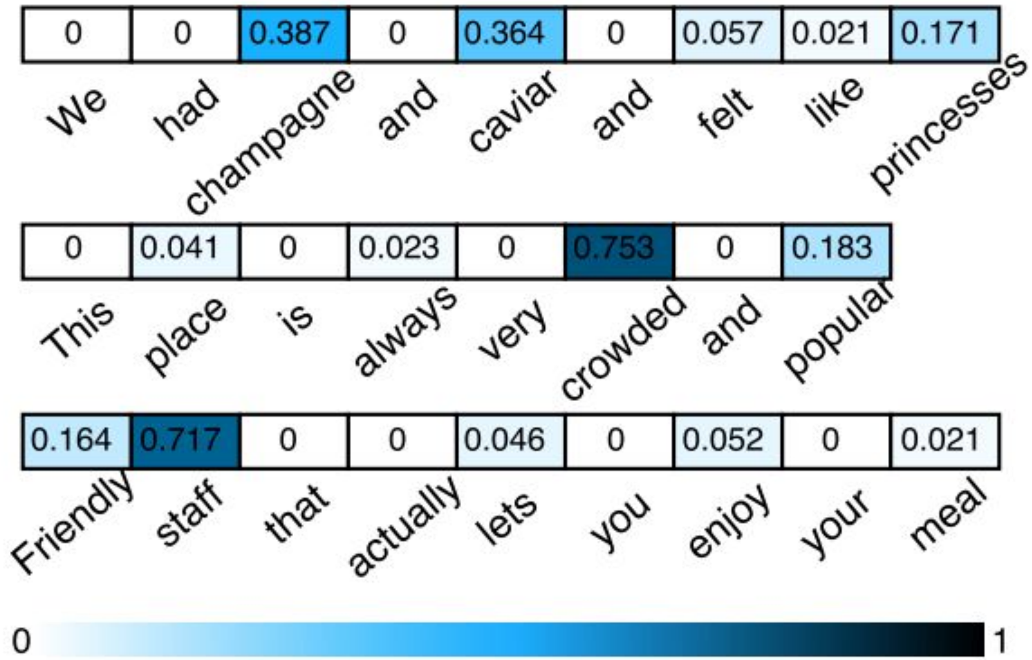


[The] service *was* *great* *!*

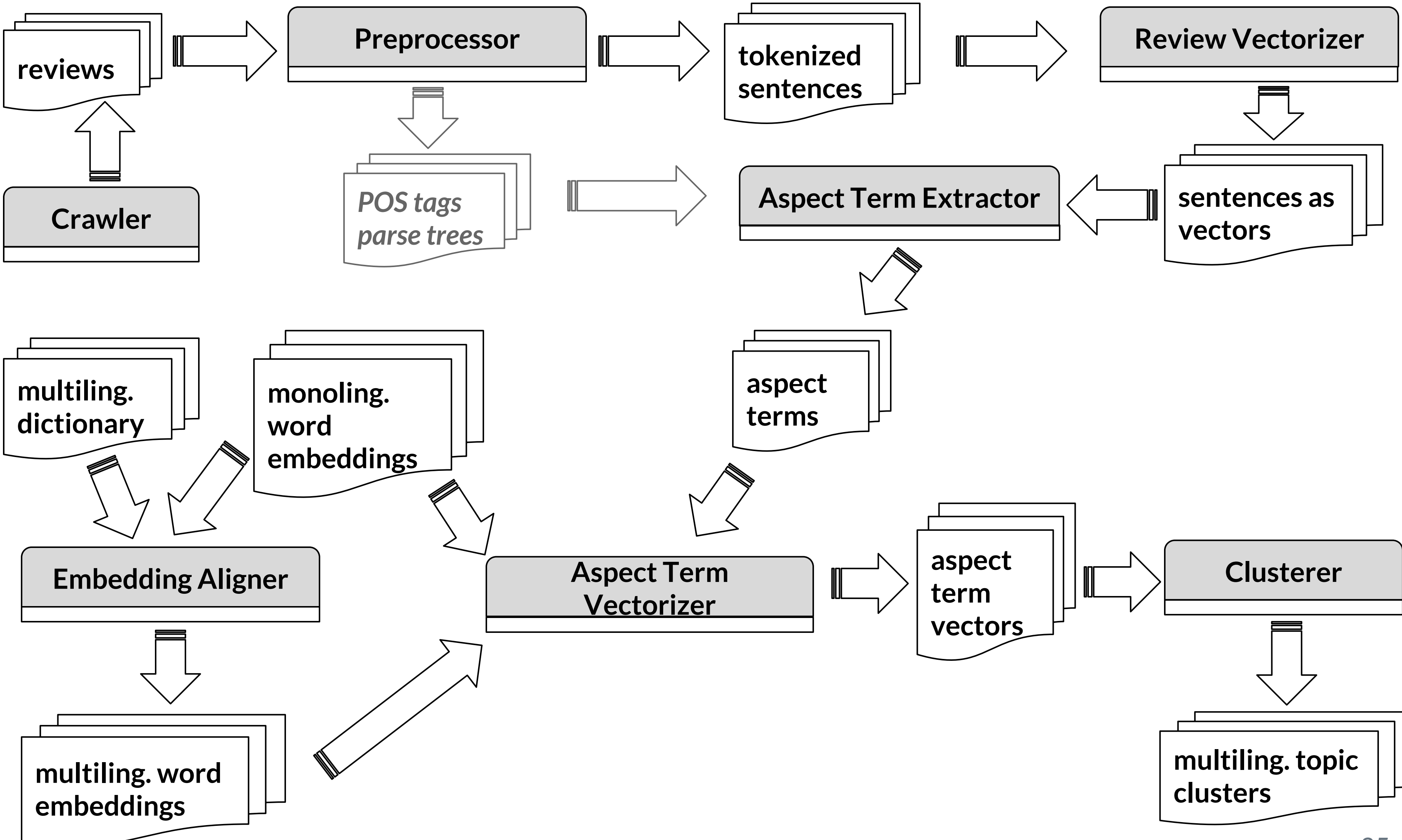
Unsupervised, attention-based

- “An Unsupervised Neural Attention Model for Aspect Extraction” (He et al., 2017)
- For each sentence, get a sentence vector
  - weighted sum of its word embeddings
  - weight: trained, capturing relevance to topics and sentence
- Recreate sentence vector from topic vectors
  - Linear combination of topic vectors and a weight vector (both trained)
- Training objective: Make recreated vector similar to sentence vector

Unsupervised, attention-based

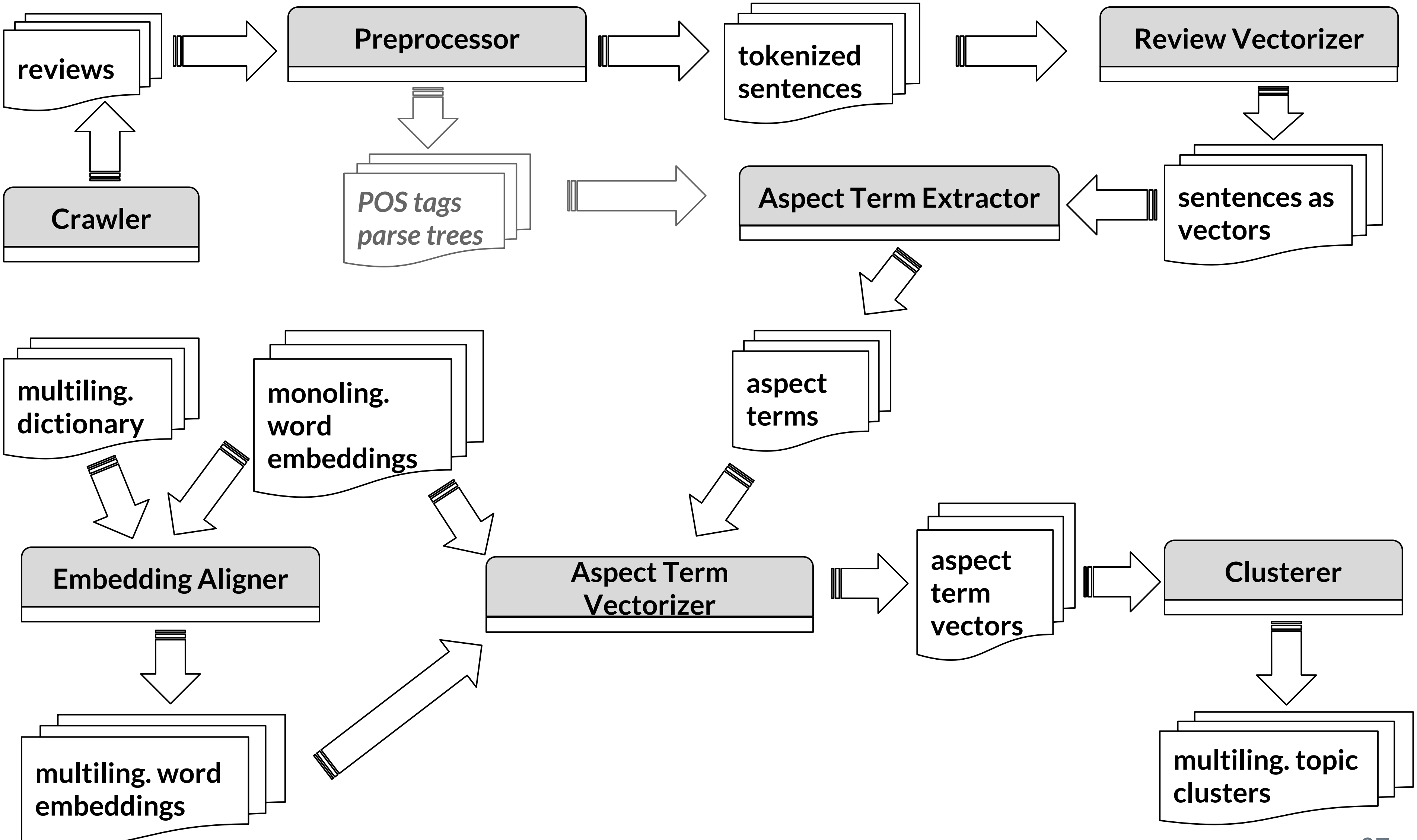


- Pick nouns with weight over threshold


reviews
Crawler
Preprocessor
POS tags
parse trees
tokenized sentences
Review Vectorizer
sentences as vectors
Aspect Term Extractor
aspect terms
multiling. dictionary
monoling. word embeddings
Embedding Aligner
multiling. word embeddings
Aspect Term Vectorizer
aspect term vectors
Clusterer
multiling. topic clusters

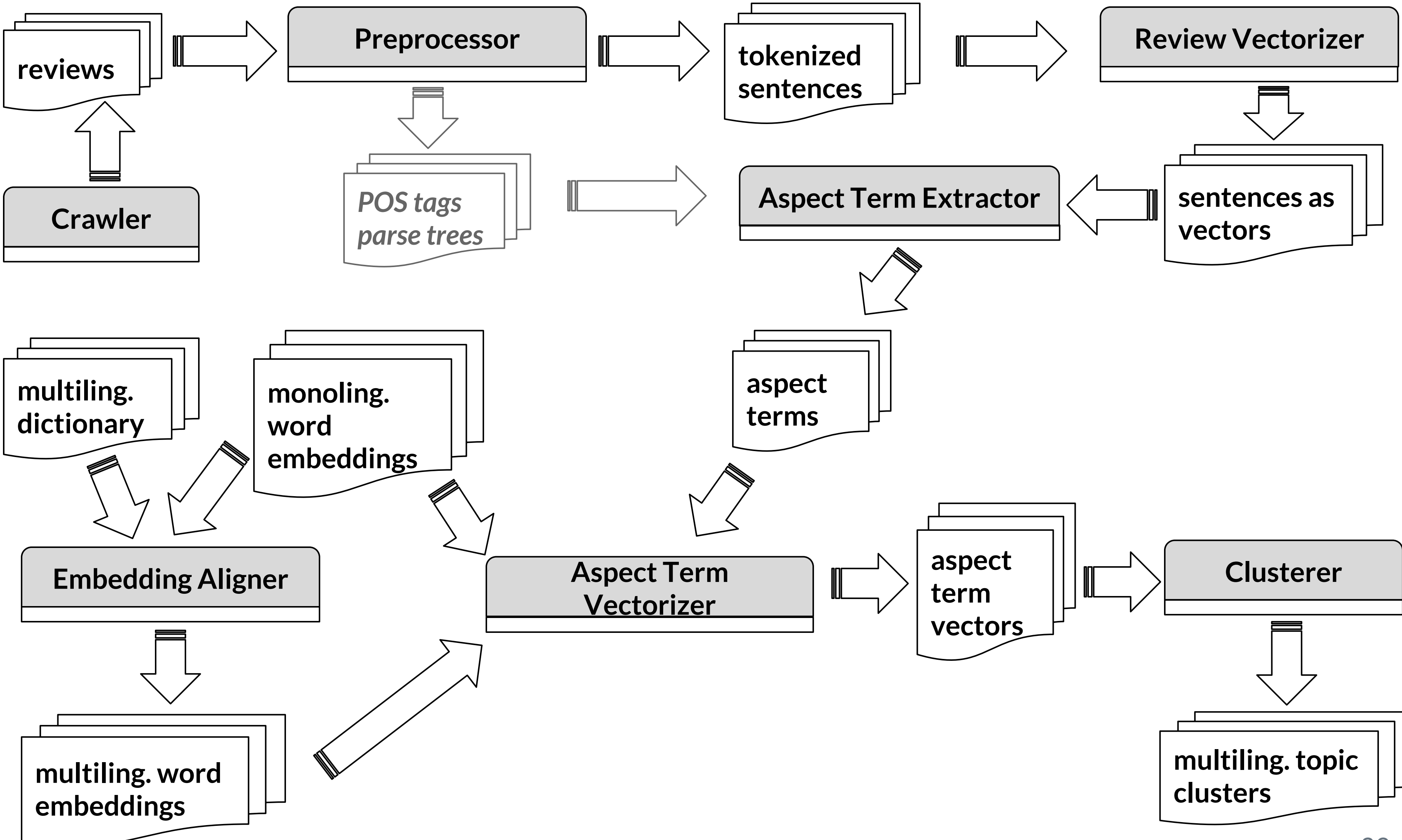
Embedding Alignment

- “Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion” (Joulin et al., 2018)
- Input: monolingual embeddings, bilingual dictionary
- Learn mapping from target to source embeddings
- Loss function: for words in dictionary, cosine similarity between vectors and their nearest neighbours
- Mapping restrained to be orthogonal (distance between embeddings is preserved)


reviews
Preprocessor
tokenized sentences
Review Vectorizer
Crawler
POS tags parse trees
Aspect Term Extractor
sentences as vectors
aspect terms
multiling. dictionary
monoling. word embeddings
Embedding Aligner
Aspect Term Vectorizer
aspect term vectors
Clusterer
multiling. word embeddings
multiling. topic clusters

Clustering

- Compare and optimize three clustering methods
  - *k*-Means
  - Affinity Propagation
  - Attention based
- Compare two ways of creating multilingual clusters
  - Directly on multilingual embeddings
  - Merge monolingual clusters
- Evaluation
  - Against coarse-grained gold standard classification
  - Homogeneity measure


reviews
Preprocessor
tokenized sentences
Review Vectorizer
Crawler
POS tags parse trees
Aspect Term Extractor
sentences as vectors
aspect terms
multiling. dictionary
monoling. word embeddings
Embedding Aligner
Aspect Term Vectorizer
aspect term vectors
Clusterer
multiling. word embeddings
multiling. topic clusters

3.

Experiments & Results

How well do the system components perform?

Aspect Extraction: Dataset

- English:
  - Annotated (SemEval 2016 Task 5)
    - 2.579 sentences (1.937 training / 642 testing)
  - Unannotated
    - 75.000 reviews, 368.551 sentences
    - From Tripadvisor, Google, OpenTable, Facebook
  - Embeddings: GloVe

Aspect Extraction: Dataset

- Finnish:
  - Annotated
    - 1.076 sentences (860 training / 216 testing)
    - Double annotated with SemEval guidelines
  - Unannotated
    - 71.730 reviews, 346.144 sentences
    - From eat.fi
  - Embeddings: FastText

AE Results: English

	Baseline	Supervised	Hybrid	Unsupervised
Precision	0.375	0.669	0.300	0.415
Recall	0.563	0.784	0.725	0.555
F1	0.450	0.722	0.424	0.473

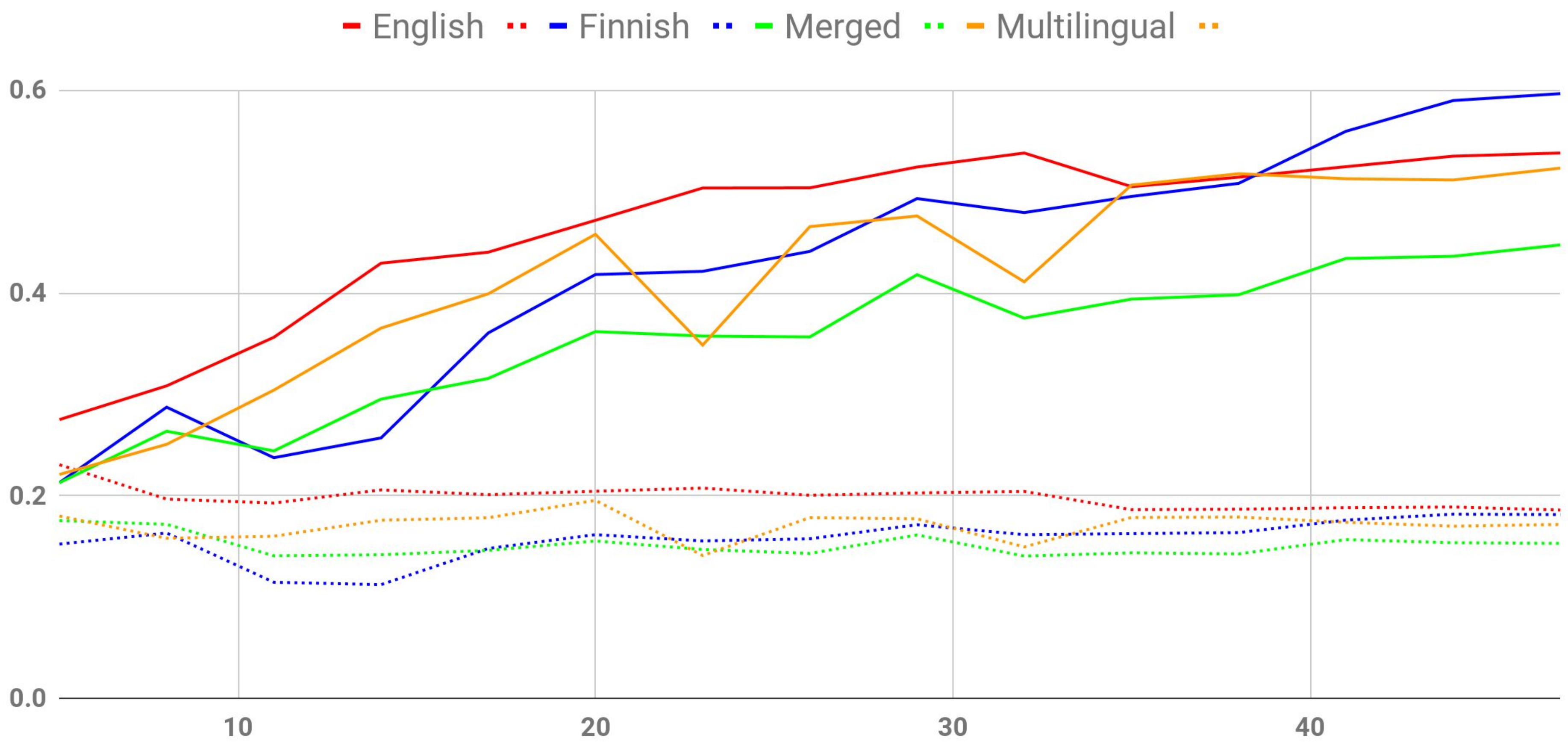
AE Results: Finnish

	Baseline	Supervised	Hybrid	Unsupervised
Precision	0.520	0.698	0.678	0.409
Recall	0.554	0.719	0.471	0.744
F1	0.537	0.707	0.556	0.528

Clustering: Dataset

- Six classes: Food, Drinks, Service, Ambiance, Location, Restaurant
- Annotated Data
  - English: 874 unique terms
  - Finnish: 623 unique terms
- Unannotated Data
  - From supervised aspect extraction method
  - 5.000 most frequent terms of each language
- Embeddings: FastText
- Maximum number of clusters: 30

Clustering Results: k-Means



Clustering Results: Comparison

	k-Means	Affinity Propagation	Attention-based
English	0.524	0.364	0.258
Finnish	0.493	0.461	0.168
Multiling. Embs.	0.475	0.344	0.023
Merged Clusters	0.418	0.295	0.076

4. Discussion

How to interpret the key findings?

Aspect Extraction

- Low performance of hybrid method
  - filtering from already low recall aspect candidates
  - training data of classifier not good enough
- Low performance of unsupervised method
  - Loss of recall from filtering outweighs gain of precision
- Supervised method
  - High performance despite exact matching requirement
  - Similar performance for Finnish and English, despite less than half training records
  - Good performance in clustering task proves performance

Clustering

- Low performance of attention-based method
  - Topic clusters are created on full reviews, not only terms
  - Clusters sometimes do not represent what we want them to (e.g. positive adverbs, first names)
- K-means works better than Affinity Propagation
  - Confirms literature that this is highly data-dependent
- Lower performance of multiling. embeddings
  - Some created clusters are still monolingual

5.

Future Work

What remains to be done?

Technical extensions

- Aligning Embeddings
  - Bigger bilingual dictionary
  - More exhaustive hyperparameter search
- Evaluation method of clustering
  - Homogeneity does not tell us which number of clusters is actually the best and does not enforce multilinguality of individual clusters
  - Subjective task, manual annotation
- Try double embeddings for clustering

Conceptional extensions

- Better network architecture for supervised extraction
  - e.g. automated search with *ENAS*
- Find better clustering method
  - k-Means has many weaknesses
  - there are many methods focused on high-dimensional data
- Analysis on entity-level
- Determine representative name for a cluster

6.

Summary

Summary

- Simple methods can work better than more complicated ones
- Tested methods designed for English can be applied to Finish with minor performance decrease
- Good results with multilingual embeddings

Thanks for listening!
Any questions?