On the Limitations of Unsupervised Bilingual Dictionary Induction

Anders Søgaard
UNIVERSITY OF COPENHAGEN

Sebastian Ruder
Insight
AYLIEN

Ivan Vulić
UNIVERSITY OF CAMBRIDGE

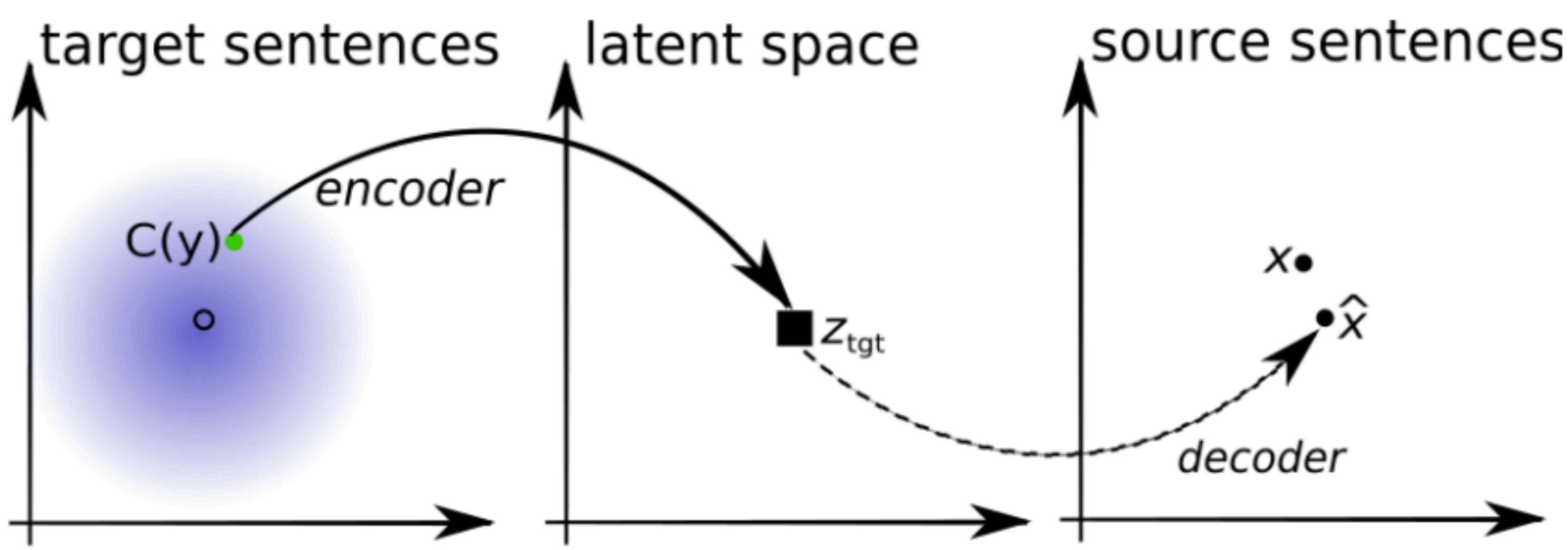
Background:
Unsupervised MT

Background: Unsupervised MT

- Recently: Unsupervised neural machine translation (Artetxe et al., ICLR 2018; Lample et al., ICLR 2018)

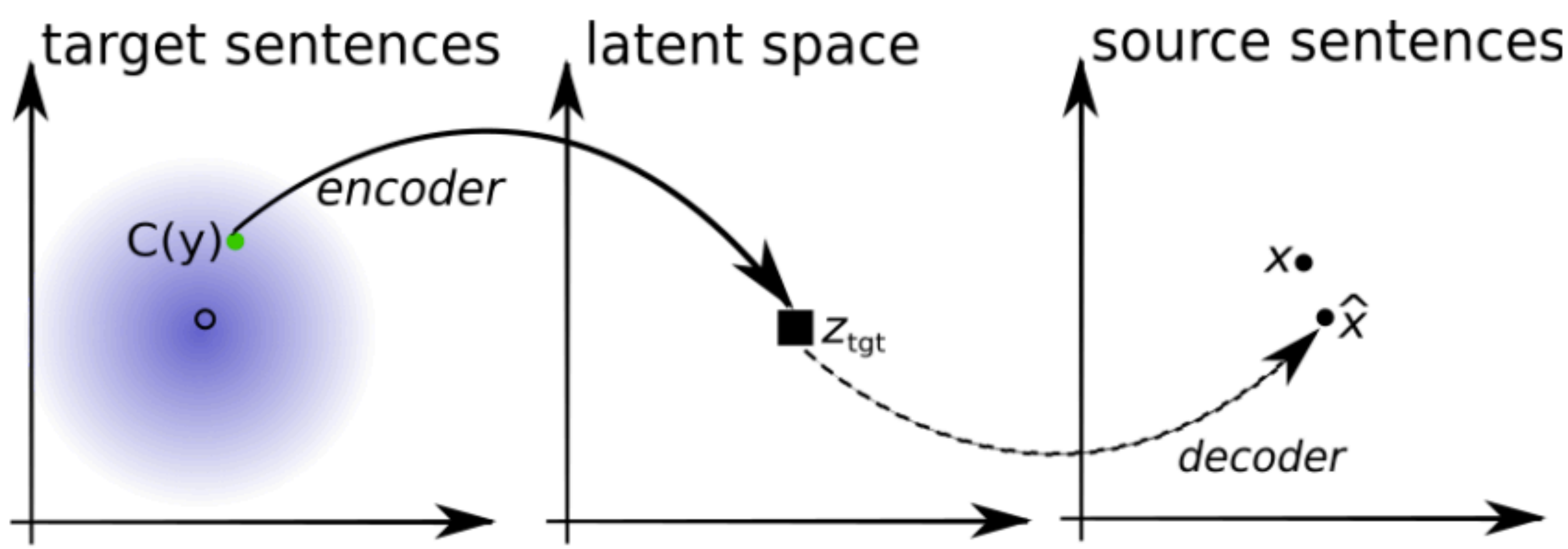
Background: Unsupervised MT

- Recently: Unsupervised neural machine translation (Artetxe et al., ICLR 2018; Lample et al., ICLR 2018)

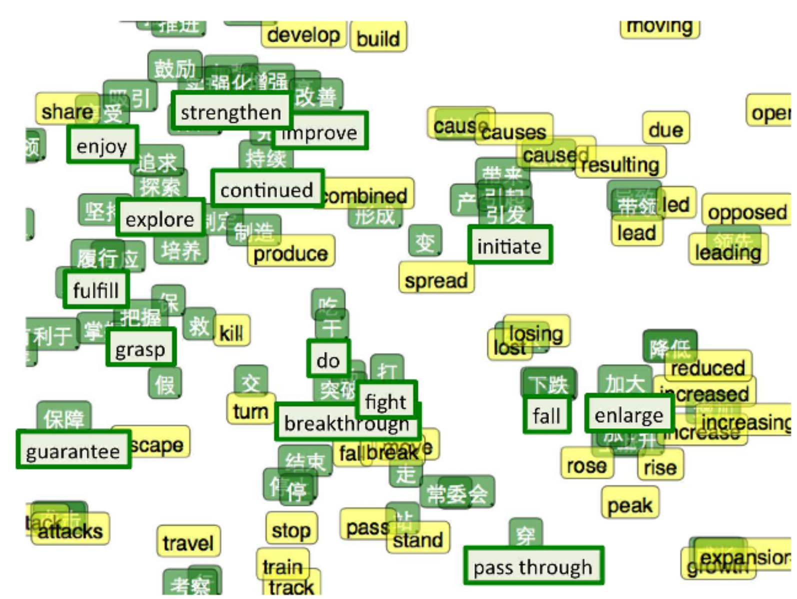


Background: Unsupervised MT

- Recently: Unsupervised neural machine translation (Artetxe et al., ICLR 2018; Lample et al., ICLR 2018)



- Key component: Initialization via unsupervised cross-lingual alignment of word embedding spaces



Background:
Cross-lingual word embeddings

Background: Cross-lingual word embeddings

- Cross-lingual word embeddings enable cross-lingual transfer

Background: Cross-lingual word embeddings

- Cross-lingual word embeddings enable cross-lingual transfer
- Most common approach: Project one word embedding space into another by learning a transformation matrix $\mathbf{W}$ between n source embeddings $\mathbf{x}_i$ and their translations $\mathbf{y}_i$

Background: Cross-lingual word embeddings

- Cross-lingual word embeddings enable cross-lingual transfer
- Most common approach: Project one word embedding space into another by learning a transformation matrix $\mathbf{W}$ between n source embeddings $\mathbf{x}_i$ and their translations $\mathbf{y}_i$

$$\sum_{i=1}^{n} \|\mathbf{W}\mathbf{x}_i - \mathbf{y}_i\|^2 \quad \text{(Mikolov et al., 2013)}$$

Background: Cross-lingual word embeddings

- Cross-lingual word embeddings enable cross-lingual transfer

- Most common approach: Project one word embedding space into another by learning a transformation matrix $\mathbf{W}$ between n source embeddings $\mathbf{x}_i$ and their translations $\mathbf{y}_i$

$$\sum_{i=1}^{n} \|\mathbf{W}\mathbf{x}_i - \mathbf{y}_i\|^2 \quad \text{(Mikolov et al., 2013)}$$

- More recently: Use an adversarial setup to learn an unsupervised mapping

Background:
Cross-lingual word embeddings

- Cross-lingual word embeddings enable cross-lingual transfer

- Most common approach: Project one word embedding space into another by learning a transformation matrix $\mathbf{W}$ between n source embeddings $\mathbf{x}_i$ and their translations $\mathbf{y}_i$

$$\sum_{i=1}^{n} \|\mathbf{W}\mathbf{x}_i - \mathbf{y}_i\|^2$$ (Mikolov et al., 2013)

- More recently: Use an adversarial setup to learn an unsupervised mapping

- Assumption: Word embedding spaces are *approximately isomorphic*, i.e. same number of vertices, connected the same way.

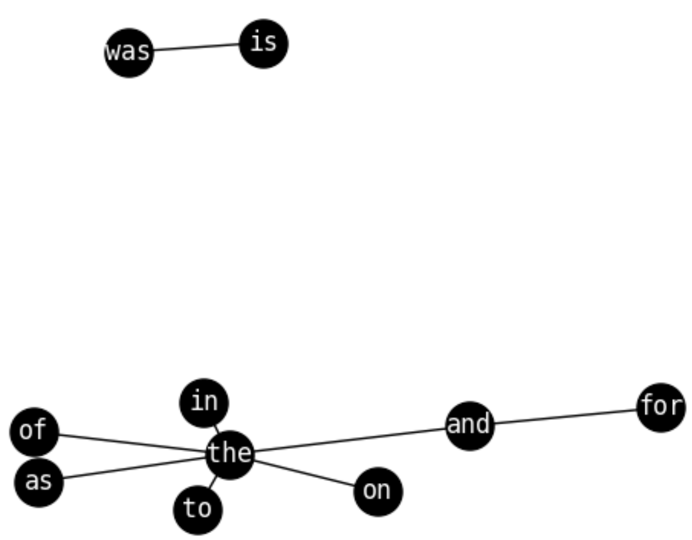
How similar are embeddings across languages?

How similar are embeddings across languages?

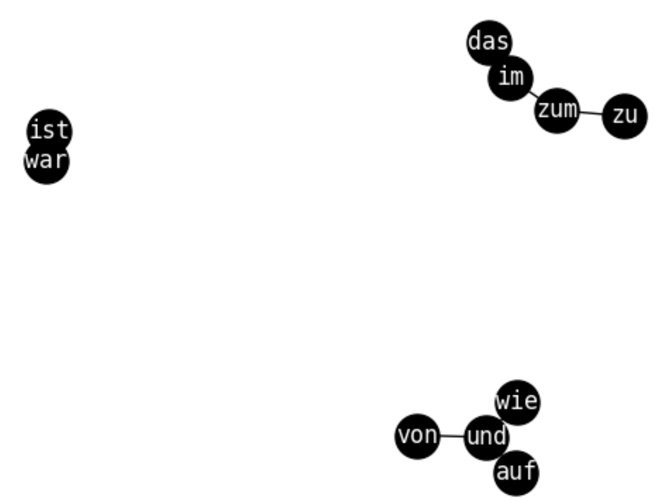
- Nearest neighbour (NN) graphs of top 10 most frequent words in English and German are not isomorphic.

How similar are embeddings across languages?

- Nearest neighbour (NN) graphs of top 10 most frequent words in English and German are not isomorphic.
- NN graphs of top 10 most frequent English words *and their translations* into German



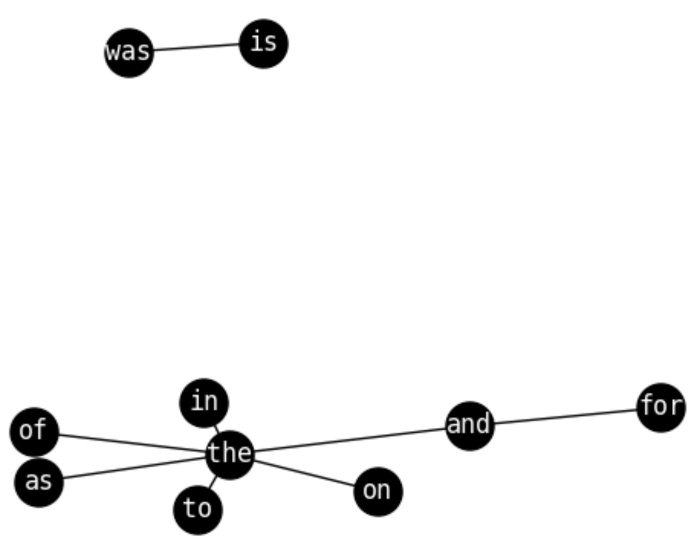
English



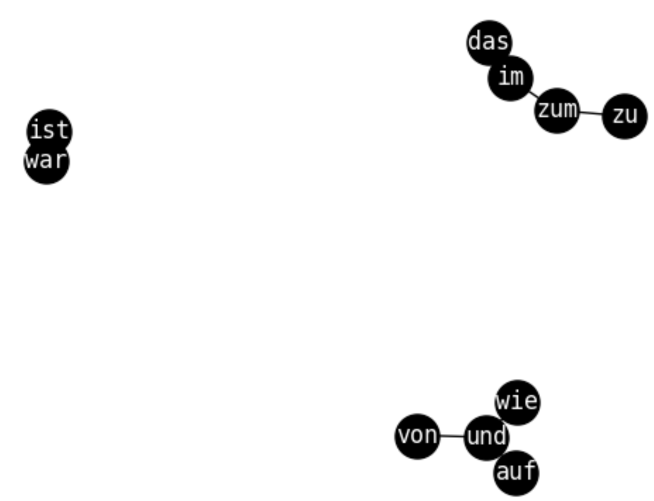
German

How similar are embeddings across languages?

- Nearest neighbour (NN) graphs of top 10 most frequent words in English and German are not isomorphic.
- NN graphs of top 10 most frequent English words *and their translations* into German



English



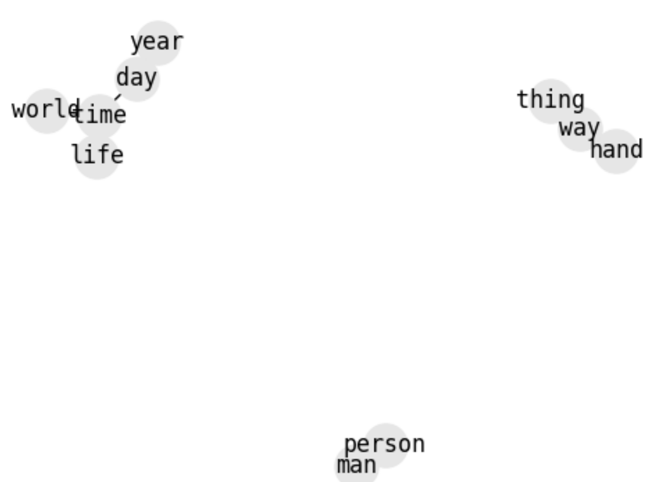
German

- Not isomorphic

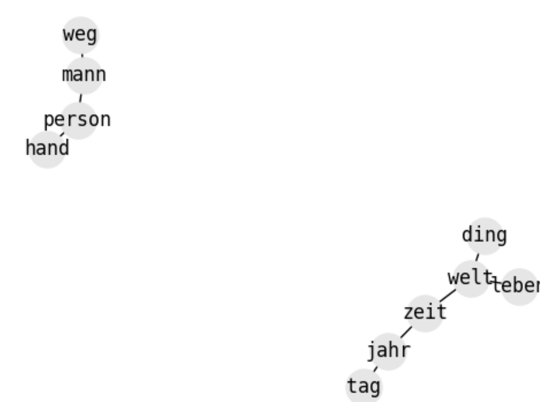
How similar are embeddings across languages?

How similar are embeddings across languages?

- NN graphs of top 10 most frequent English *nouns* and their translations



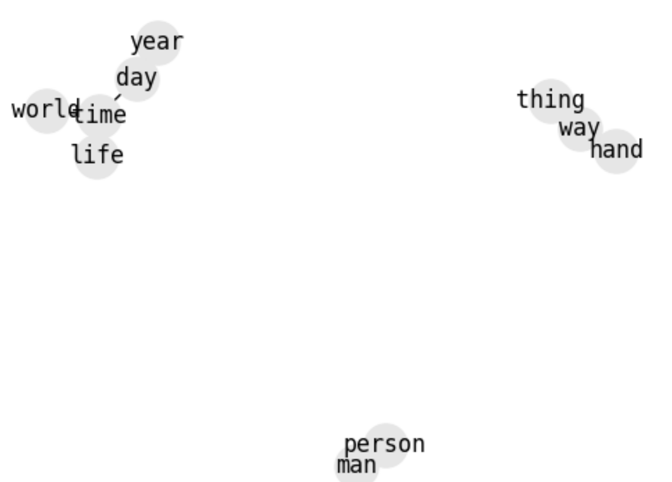
English



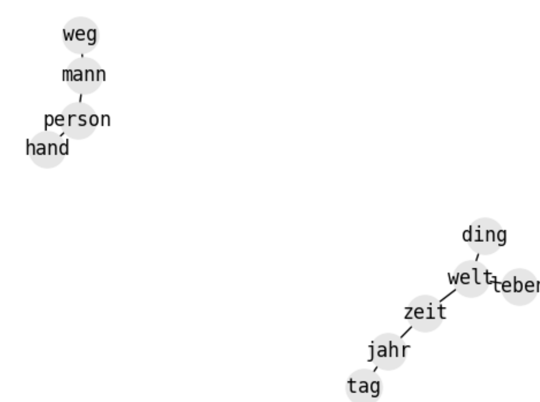
German

How similar are embeddings across languages?

- NN graphs of top 10 most frequent English *nouns* and their translations



English

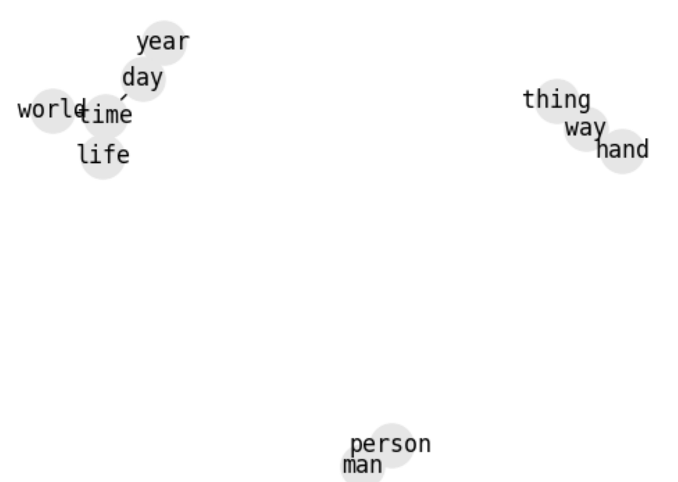


German

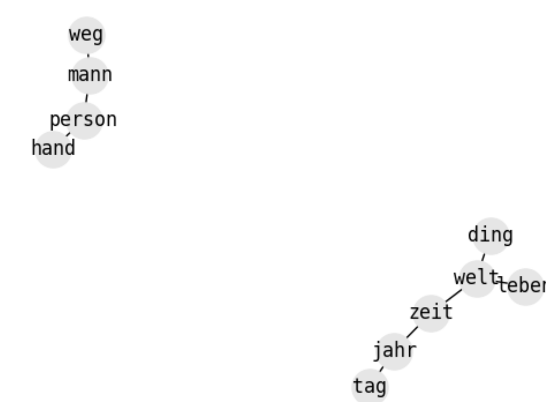
- Not isomorphic

How similar are embeddings across languages?

- NN graphs of top 10 most frequent English *nouns* and their translations



English



German

- Not isomorphic

Word embeddings are *not* approximately isomorphic across languages.

How do we quantify similarity?

How do we quantify similarity?

- Need a metric to measure how similar two NN graphs G_1 and G_2 of different languages are

How do we quantify similarity?

- Need a metric to measure how similar two NN graphs G_1 and G_2 of different languages are
- Propose **eigenvector similarity**

How do we quantify similarity?

- Need a metric to measure how similar two NN graphs G_1 and G_2 of different languages are
- Propose **eigenvector similarity**
- A_1, A_2 : adjacency matrices of G_1, G_2

How do we quantify similarity?

- Need a metric to measure how similar two NN graphs G_1 and G_2 of different languages are
- Propose **eigenvector similarity**
- A_1, A_2 : adjacency matrices of G_1, G_2
- D_1, D_2 : degree matrices of G_1, G_2

How do we quantify similarity?

- Need a metric to measure how similar two NN graphs G_1 and G_2 of different languages are
- Propose **eigenvector similarity**
- A_1, A_2 : adjacency matrices of G_1, G_2
- D_1, D_2: degree matrices of G_1, G_2
- $L_1 = D_1 - A_1, L_2 = D_2 - A_2$: Laplacians of G_1, G_2

How do we quantify similarity?

- Need a metric to measure how similar two NN graphs G_1 and G_2 of different languages are
- Propose **eigenvector similarity**
- A_1, A_2 : adjacency matrices of G_1, G_2
- D_1, D_2 : degree matrices of G_1, G_2
- $L_1 = D_1 - A_1, L_2 = D_2 - A_2$: Laplacians of G_1, G_2
- λ_1, λ_2 : eigenvalues (spectra) of L_1, L_2

How do we quantify similarity?

- Need a metric to measure how similar two NN graphs G_1 and G_2 of different languages are
- Propose **eigenvector similarity**
- A_1, A_2 : adjacency matrices of G_1, G_2
- D_1, D_2 : degree matrices of G_1, G_2
- $L_1 = D_1 - A_1, L_2 = D_2 - A_2$: Laplacians of G_1, G_2
- λ_1, λ_2 : eigenvalues (spectra) of L_1, L_2
- Metric: $\Delta = \sum_{i=1}^{k} (\lambda_{1_i} - \lambda_{2_i})^2$ where $k = \min_j \{ \frac{\sum_{i=1}^{k} \lambda_{j_i}}{\sum_{i=1}^{n} \lambda_{j_i}} > 0.9 \}$

How do we quantify similarity?

- Metric: $\Delta = \sum_{i=1}^{k} (\lambda_{1_i} - \lambda_{2_i})^2$ where $k = \min_j \{ \frac{\sum_{i=1}^{k} \lambda_{j_i}}{\sum_{i=1}^{n} \lambda_{j_i}} > 0.9 \}$

How do we quantify similarity?

- Quantifies how much two NN graphs are isospectral, i.e. they have the same spectrum (same sets of eigenvalues).

- Metric: $\Delta = \sum_{i=1}^{k} (\lambda_{1_i} - \lambda_{2_i})^2$ where $k = \min_j \{ \frac{\sum_{i=1}^{k} \lambda_{j_i}}{\sum_{i=1}^{n} \lambda_{j_i}} > 0.9 \}$

How do we quantify similarity?

- Quantifies how much two NN graphs are isospectral, i.e. they have the same spectrum (same sets of eigenvalues).
- Isomorphic $\rightarrow$ isospectral, but isospectral $\nrightarrow$ isomorphic

- Metric: $\Delta = \sum_{i=1}^{k} (\lambda_{1_i} - \lambda_{2_i})^2$ where $k = \min_j \{ \frac{\sum_{i=1}^{k} \lambda_{j_i}}{\sum_{i=1}^{n} \lambda_{j_i}} > 0.9 \}$

How do we quantify similarity?

- Quantifies how much two NN graphs are isospectral, i.e. they have the same spectrum (same sets of eigenvalues).
- Isomorphic $\rightarrow$ isospectral, but isospectral $\nrightarrow$ isomorphic
- $\Delta : G_1, G_2 \rightarrow [0,\infty)$

- Metric: $\Delta = \sum_{i=1}^{k} (\lambda_{1_i} - \lambda_{2_i})^2$ where $k = \min_j \{ \frac{\sum_{i=1}^{k} \lambda_{j_i}}{\sum_{i=1}^{n} \lambda_{j_i}} > 0.9 \}$

How do we quantify similarity?

- Quantifies how much two NN graphs are isospectral, i.e. they have the same spectrum (same sets of eigenvalues).
- Isomorphic $\rightarrow$ isospectral, but isospectral $\nrightarrow$ isomorphic
- $\Delta : G_1, G_2 \rightarrow [0,\infty)$
- $\Delta = 0$: G_1, G_2 are isospectral (very similar)

- Metric: $\Delta = \sum_{i=1}^{k} (\lambda_{1_i} - \lambda_{2_i})^2$ where $k = \min_j \{ \frac{\sum_{i=1}^{k} \lambda_{j_i}}{\sum_{i=1}^{n} \lambda_{j_i}} > 0.9 \}$

How do we quantify similarity?

- Quantifies how much two NN graphs are isospectral, i.e. they have the same spectrum (same sets of eigenvalues).
- Isomorphic $\rightarrow$ isospectral, but isospectral $\nrightarrow$ isomorphic
- $\Delta : G_1, G_2 \rightarrow [0,\infty)$
- $\Delta = 0$: G_1, G_2 are isospectral (very similar)
- $\Delta \rightarrow \infty$: G_1, G_2 become less similar

- Metric: $\Delta = \sum_{i=1}^{k} (\lambda_{1_i} - \lambda_{2_i})^2$ where $k = \min_j \{ \frac{\sum_{i=1}^{k} \lambda_{j_i}}{\sum_{i=1}^{n} \lambda_{j_i}} > 0.9 \}$

Unsupervised cross-lingual learning assumptions

Unsupervised cross-lingual learning assumptions

- Besides isomorphism, several other implicit assumptions

Unsupervised cross-lingual learning assumptions

- Besides isomorphism, several other implicit assumptions
- May or may not scale to low-resource languages

Unsupervised cross-lingual learning assumptions

- Besides isomorphism, several other implicit assumptions
- May or may not scale to low-resource languages

	Conneau et al. (2018)	This work

Unsupervised cross-lingual learning assumptions

- Besides isomorphism, several other implicit assumptions
- May or may not scale to low-resource languages

	Conneau et al. (2018)	This work
Languages	Dependent-marking, fusional and isolating	Agglutinative, many cases

Unsupervised cross-lingual learning assumptions

- Besides isomorphism, several other implicit assumptions
- May or may not scale to low-resource languages

	Conneau et al. (2018)	This work
Languages	Dependent-marking, fusional and isolating	Agglutinative, many cases
Corpora	Comparable (Wikipedia)	Different domains

Unsupervised cross-lingual learning assumptions

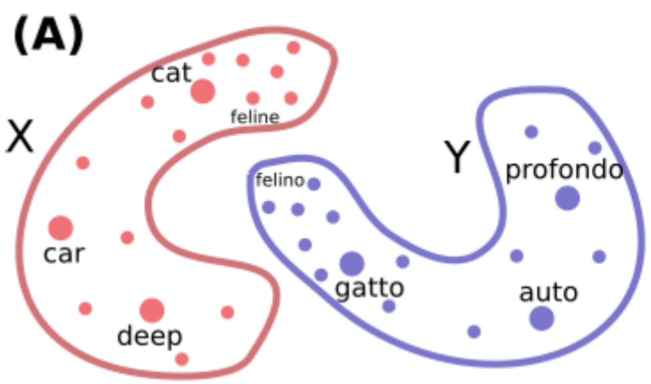
- Besides isomorphism, several other implicit assumptions
- May or may not scale to low-resource languages

	Conneau et al. (2018)	This work
Languages	Dependent-marking, fusional and isolating	Agglutinative, many cases
Corpora	Comparable (Wikipedia)	Different domains
Algorithms/ hyperparameters	Same	Different

Conneau et al. (2018)

Conneau et al. (2018)

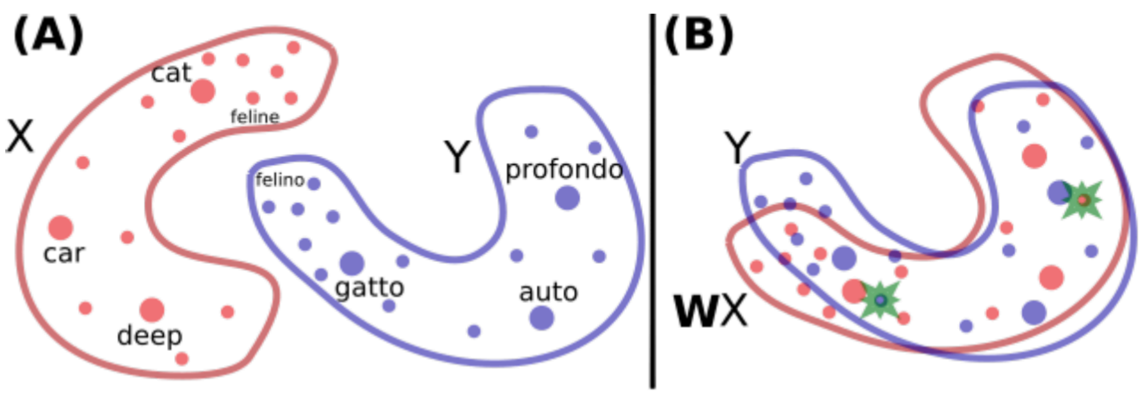
1. **Monolingual word embeddings:**
   Learn monolingual vector spaces X and Y.



Conneau et al. (2018)

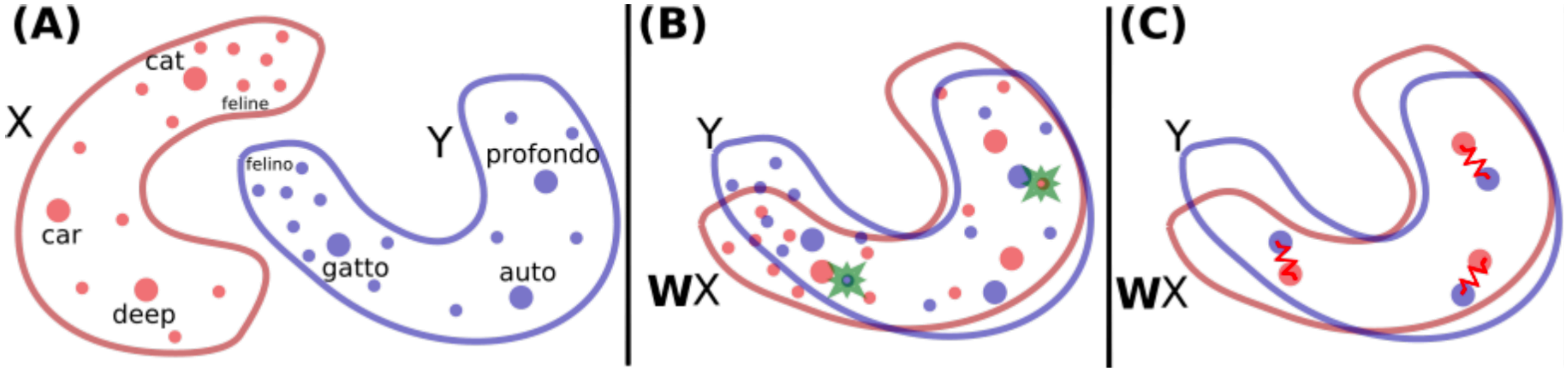
1. **Monolingual word embeddings:**
   Learn monolingual vector spaces X and Y.

2. **Adversarial mapping:**
   Learn a translation matrix W. Train discriminator to discriminate samples from WX and Y.



Conneau et al. (2018)

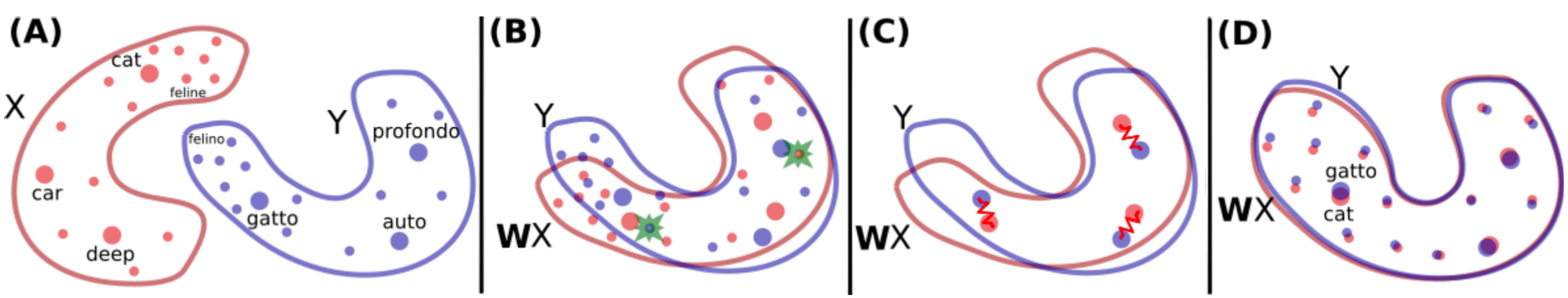
3. **Refinement (Procrustes analysis):**
   Build bilingual dictionary of frequent words using W. Learn a new W based on frequent word pairs.



Conneau et al. (2018)

3. **Refinement (Procrustes analysis):**
   Build bilingual dictionary of frequent words using W. Learn a new W based on frequent word pairs.

4. **Cross-domain similarity local scaling (CSLS):**
   Use similarity measure that increases similarity of isolated word vectors, decreases similarity of vectors in dense areas.



A simple weakly supervised method

A simple weakly supervised method

- Extract identically spelled words in both languages

A simple weakly supervised method

- Extract identically spelled words in both languages
- Use these as bilingual seed words

A simple weakly supervised method

- Extract identically spelled words in both languages
- Use these as bilingual seed words
- Run refinement step of Conneau et al. (2018)

Experiments:
Bilingual dictionary induction

Experiments: Bilingual dictionary induction

- Given a list of source language words, find the closest target language word in the cross-lingual embedding space

Experiments: Bilingual dictionary induction

- Given a list of source language words, find the closest target language word in the cross-lingual embedding space
- Compare against a gold standard dictionary

Experiments: Bilingual dictionary induction

- Given a list of source language words, find the closest target language word in the cross-lingual embedding space
- Compare against a gold standard dictionary
- Metric: Precision at 1 (P@1)

Experiments:
Bilingual dictionary induction

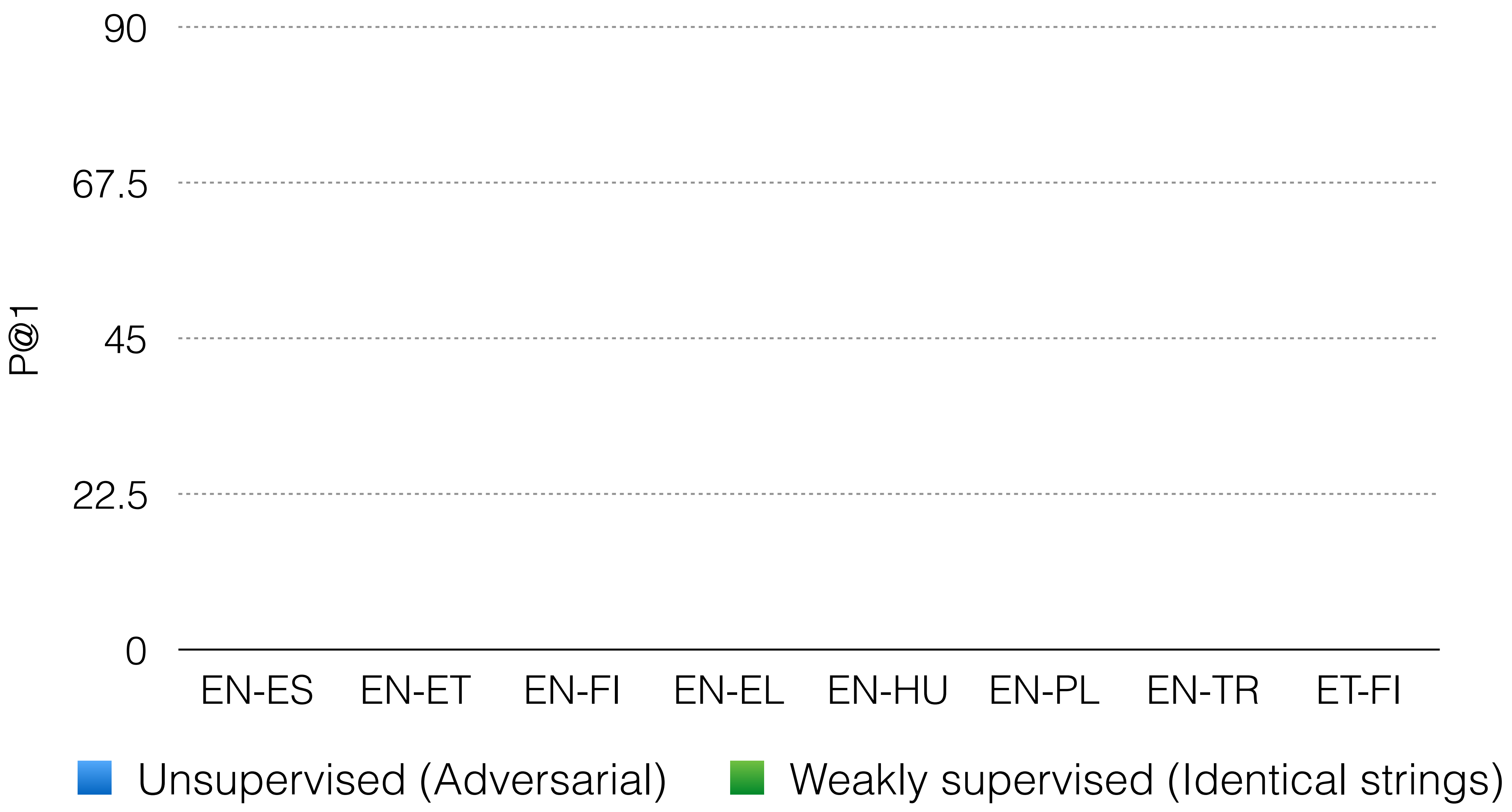
- Given a list of source language words, find the closest target language word in the cross-lingual embedding space
- Compare against a gold standard dictionary
- Metric: Precision at 1 (P@1)
- Use fastText monolingual embeddings

Experiments:
Bilingual dictionary induction

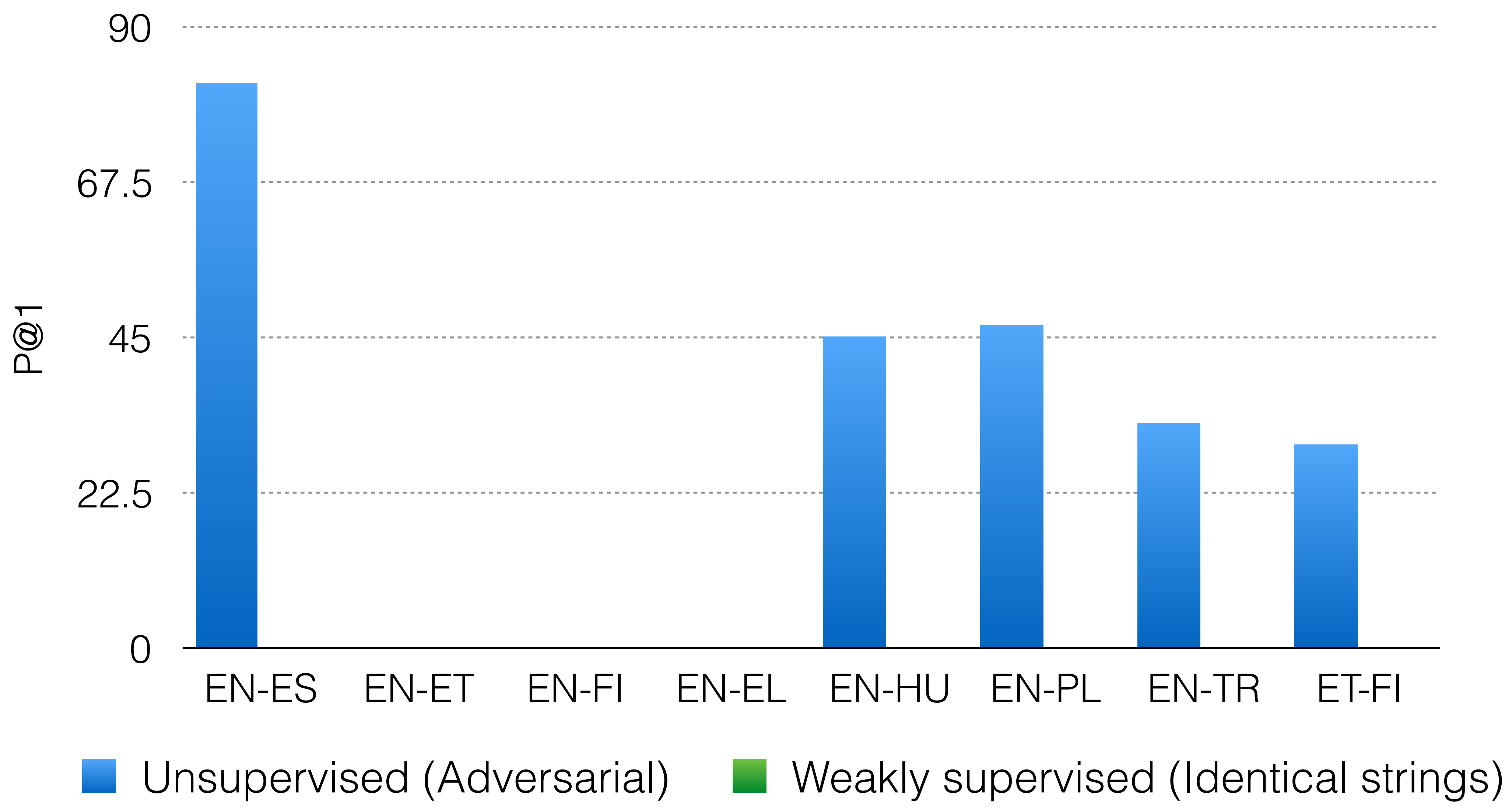
- Given a list of source language words, find the closest target language word in the cross-lingual embedding space
- Compare against a gold standard dictionary
- Metric: Precision at 1 (P@1)
- Use fastText monolingual embeddings

	Conneau et al. (2018)	This work
Languages (English to)	French, German, Chinese, Russian, Spanish	Estonian (ET), Finnish (FI), Greek (EL), Hungarian (HU), Polish (PL), Turkish

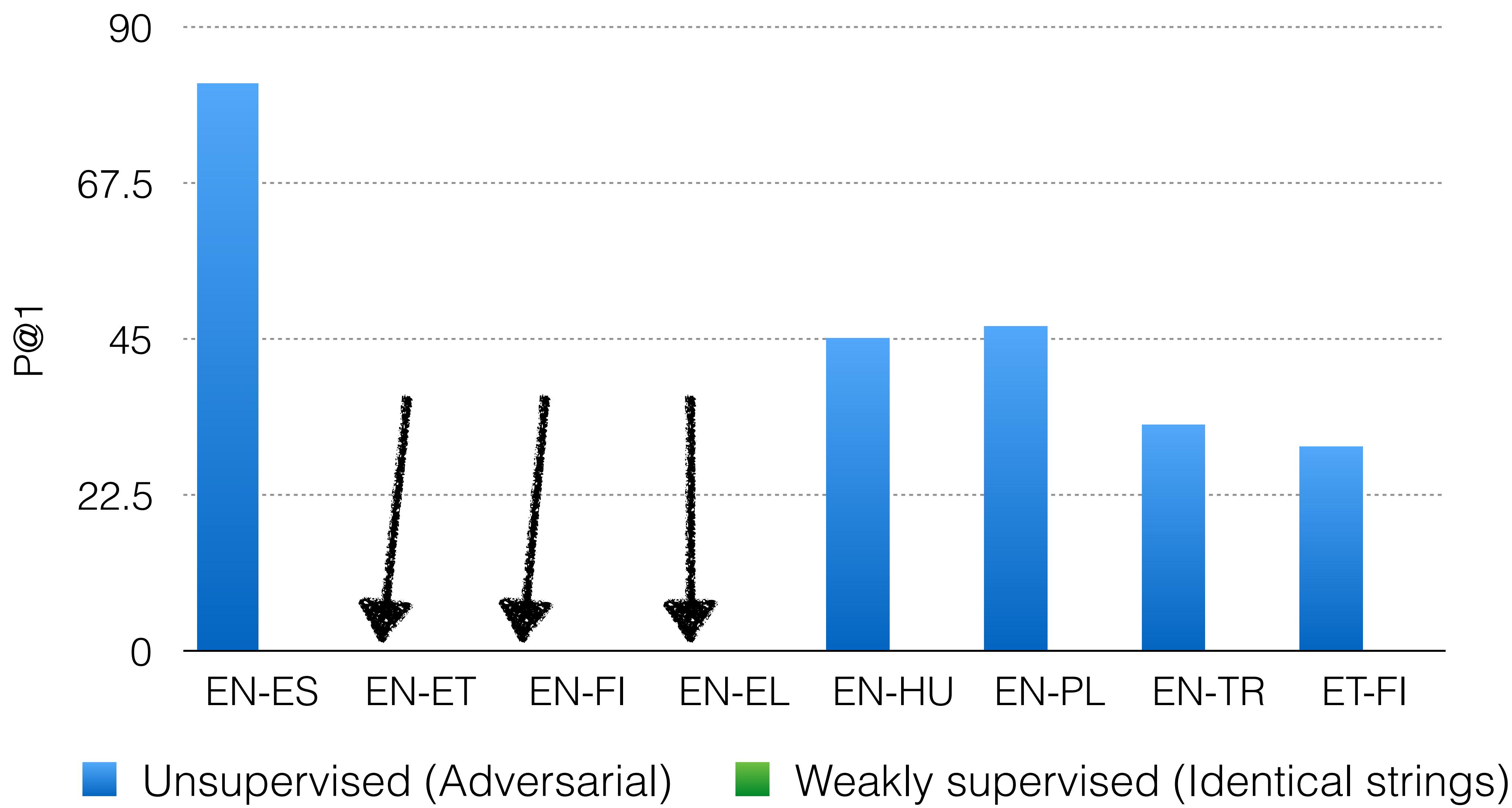
Impact of language similarity



Impact of language similarity

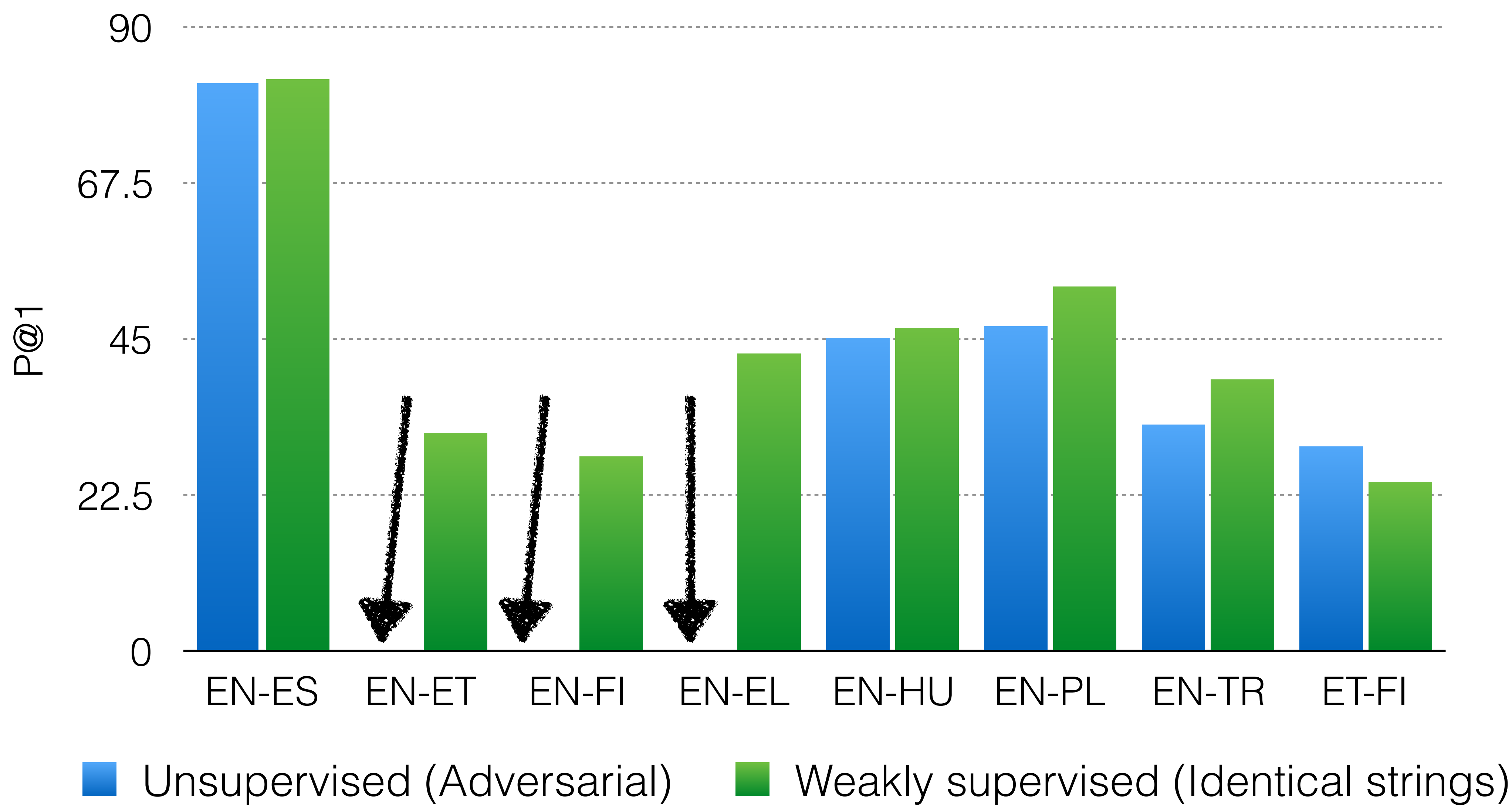


Impact of language similarity



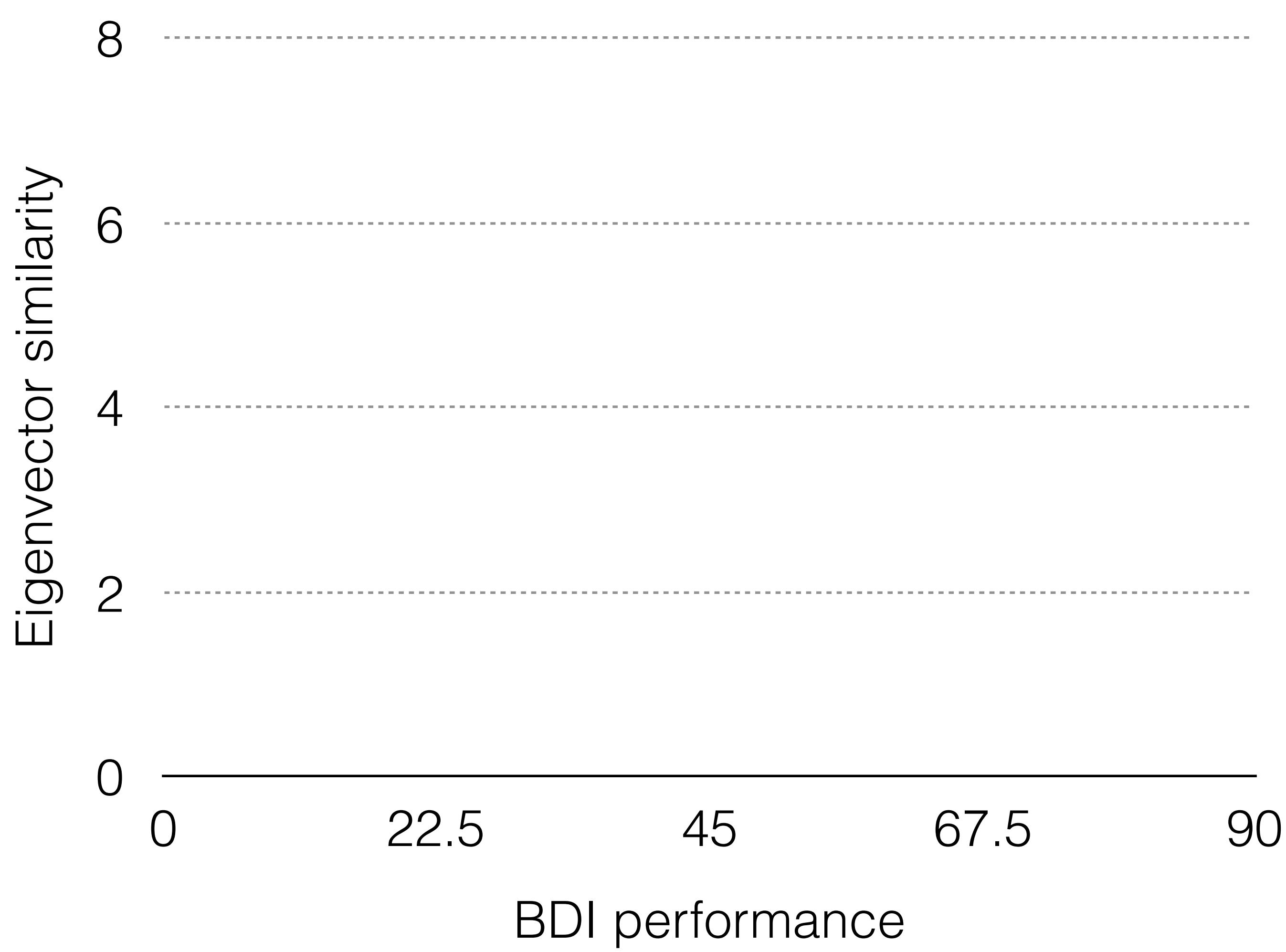
- Unsupervised approaches are challenged by languages that are not isolating and not dependent marking

Impact of language similarity

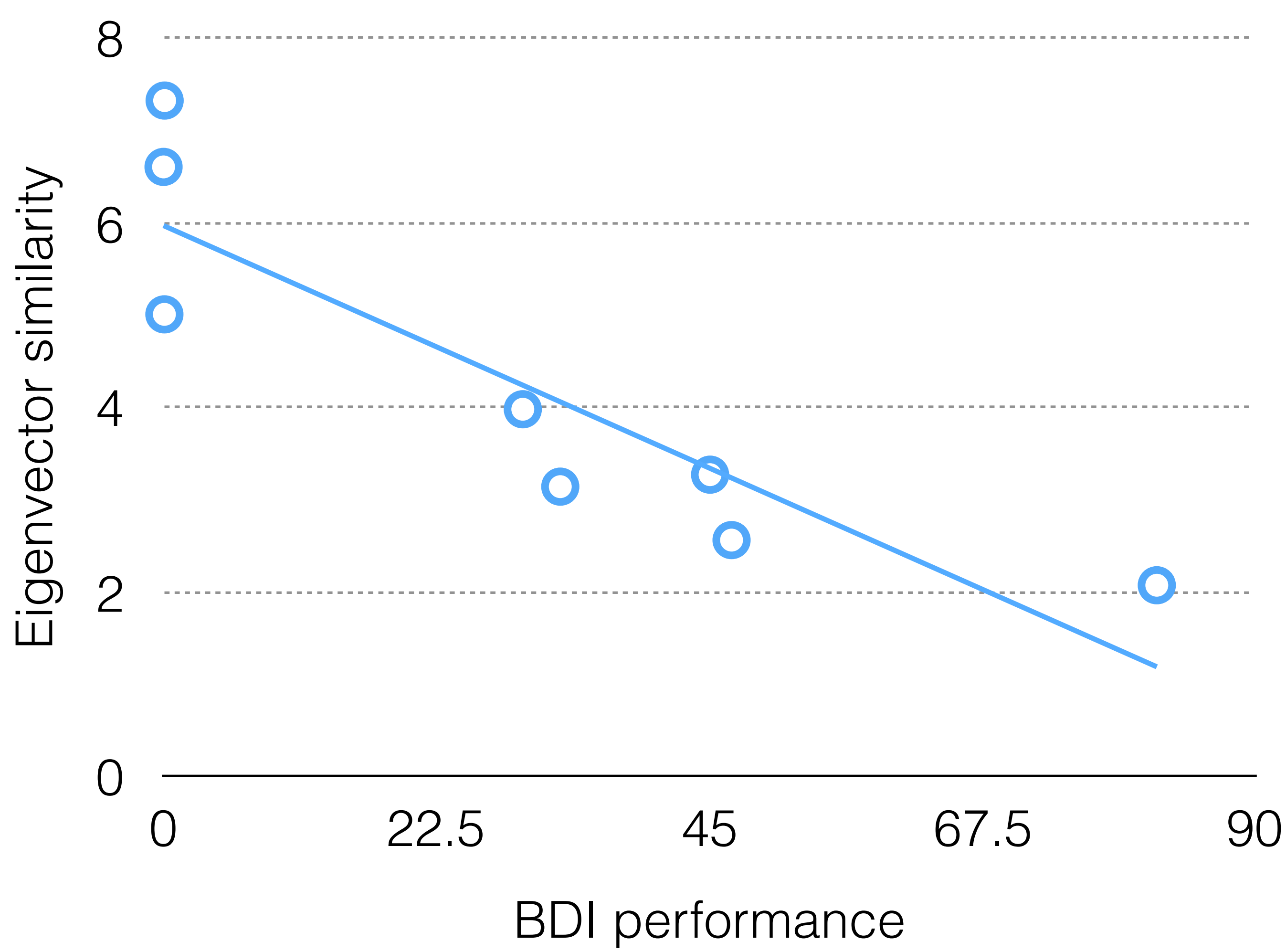


- Unsupervised approaches are challenged by languages that are not isolating and not dependent marking
- Naive supervision leads to competitive performance on similar language pairs and better results for dissimilar pairs

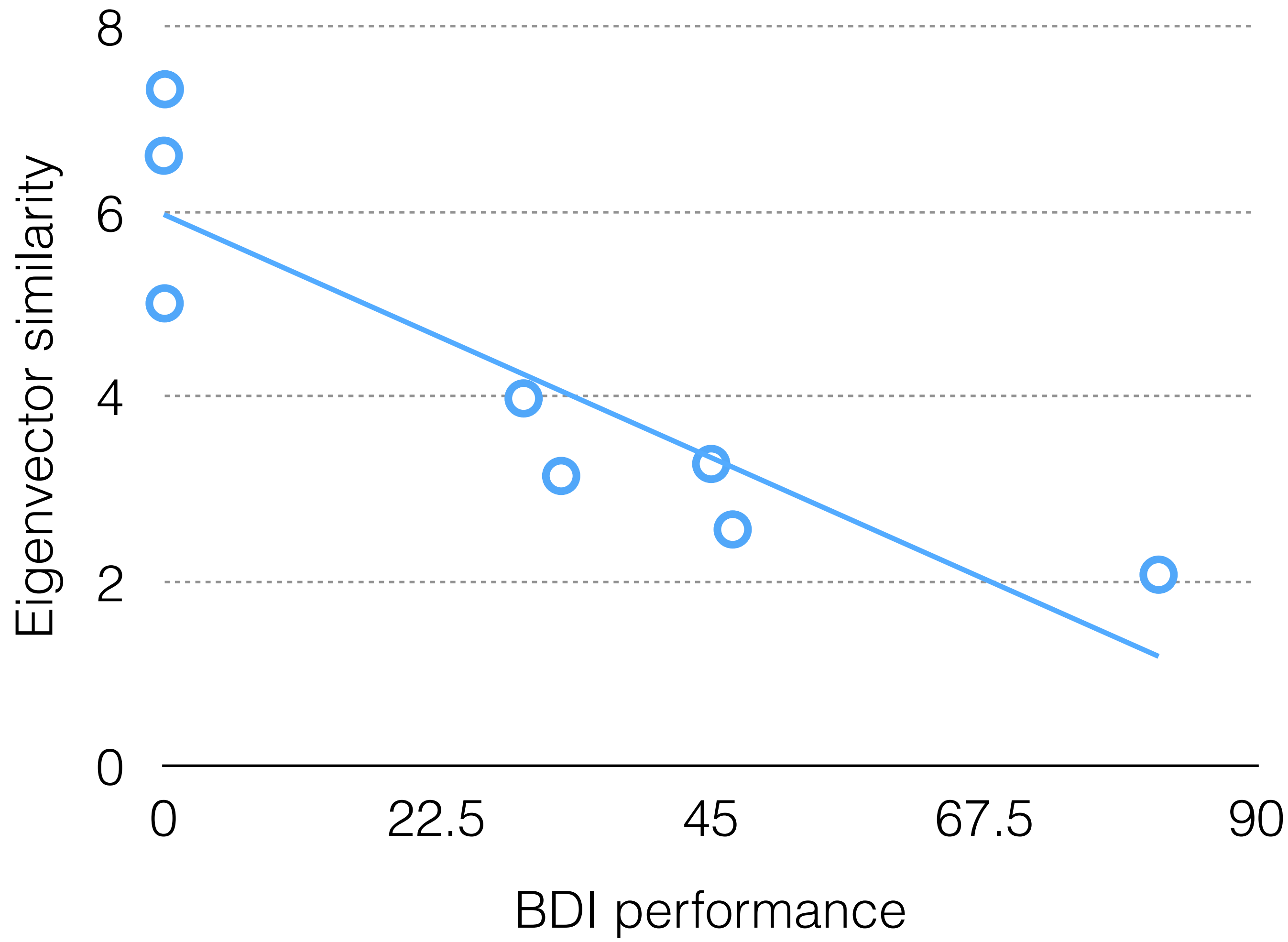
Impact of language similarity



Impact of language similarity



Impact of language similarity



- Eigenvector similarity strongly correlates with BDI performance ($\rho \sim 0.89$)

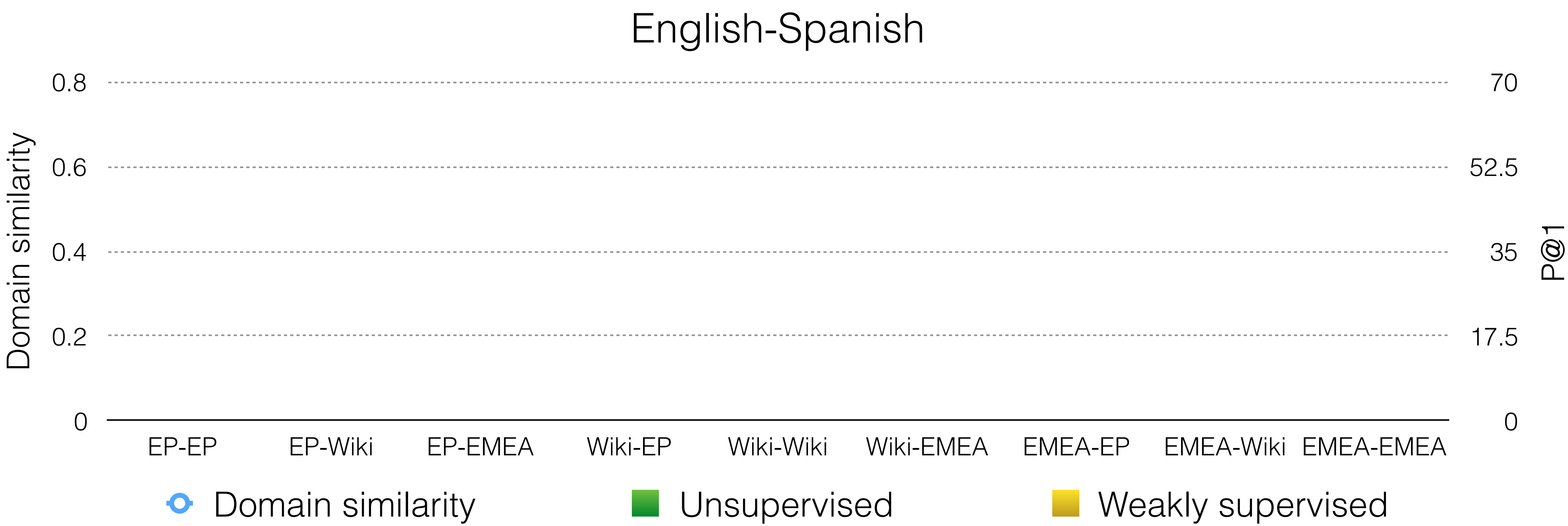
Impact of domain differences

Impact of domain differences

- Source and target embeddings induced on 3 corpora: EuroParl (EP), Wikipedia (Wiki), Medical (EMEA)

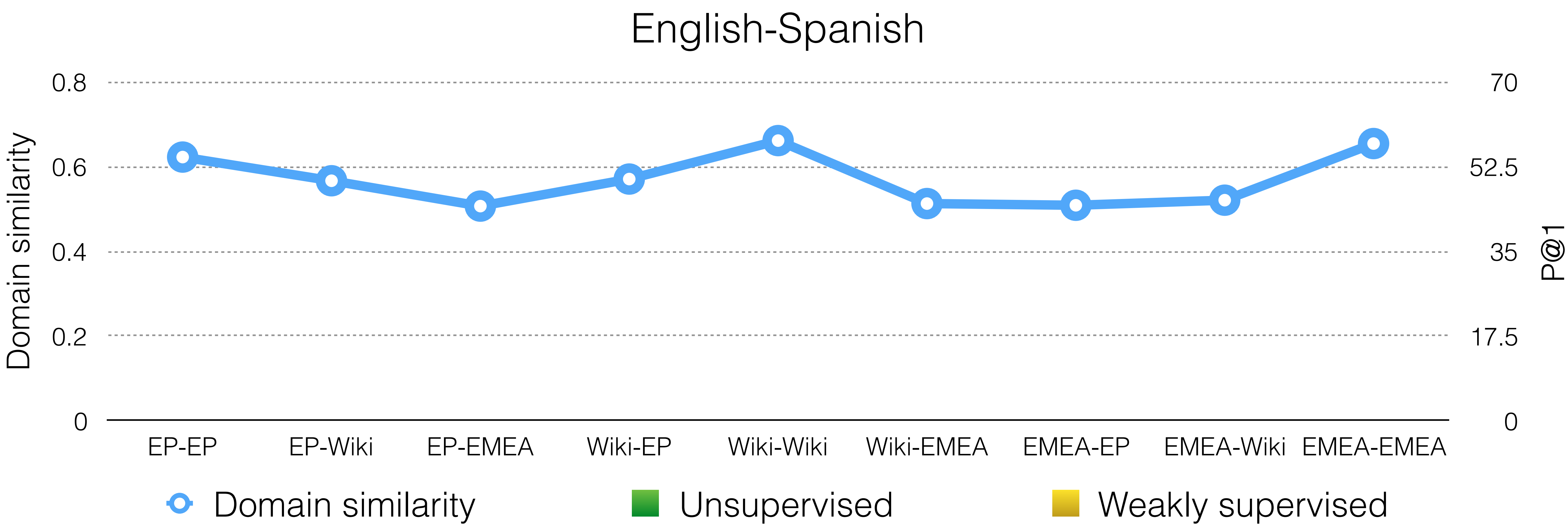
Impact of domain differences

- Source and target embeddings induced on 3 corpora: EuroParl (EP), Wikipedia (Wiki), Medical (EMEA)



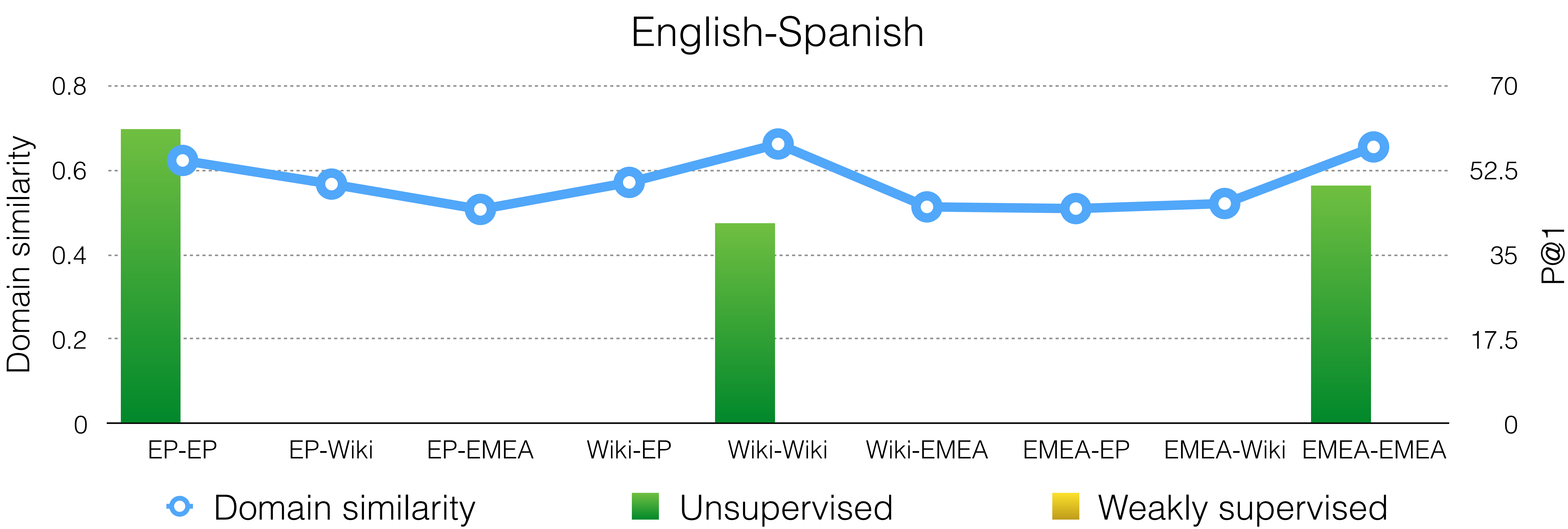
Impact of domain differences

- Source and target embeddings induced on 3 corpora: EuroParl (EP), Wikipedia (Wiki), Medical (EMEA)



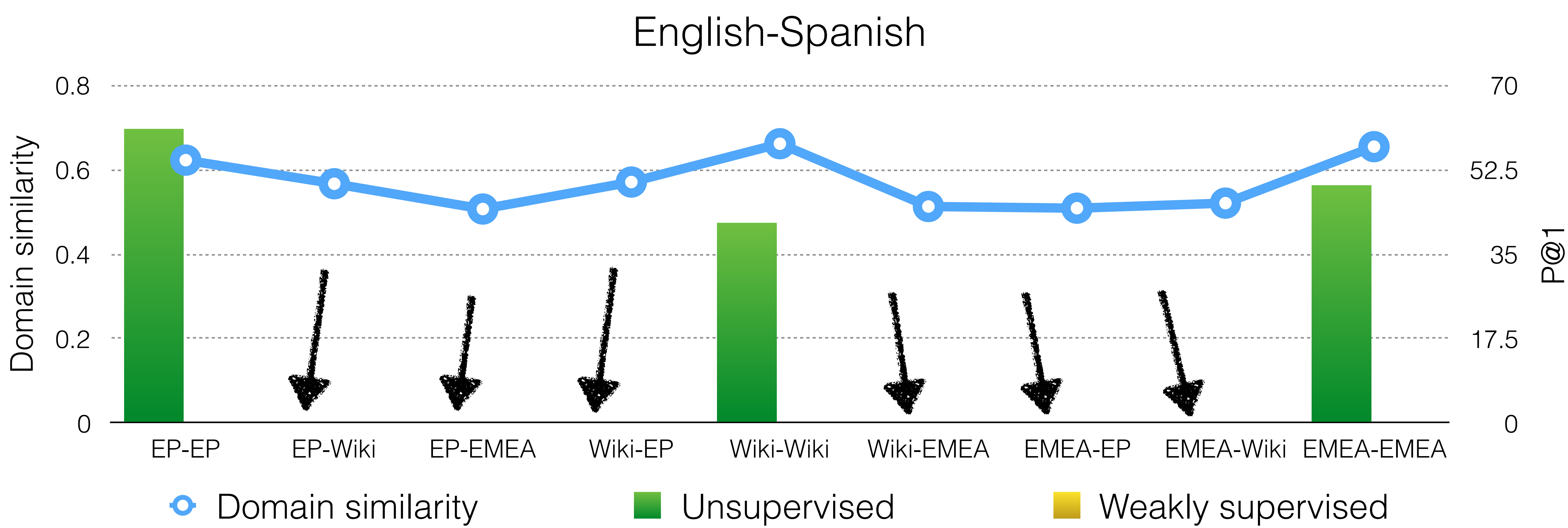
Impact of domain differences

- Source and target embeddings induced on 3 corpora: EuroParl (EP), Wikipedia (Wiki), Medical (EMEA)



Impact of domain differences

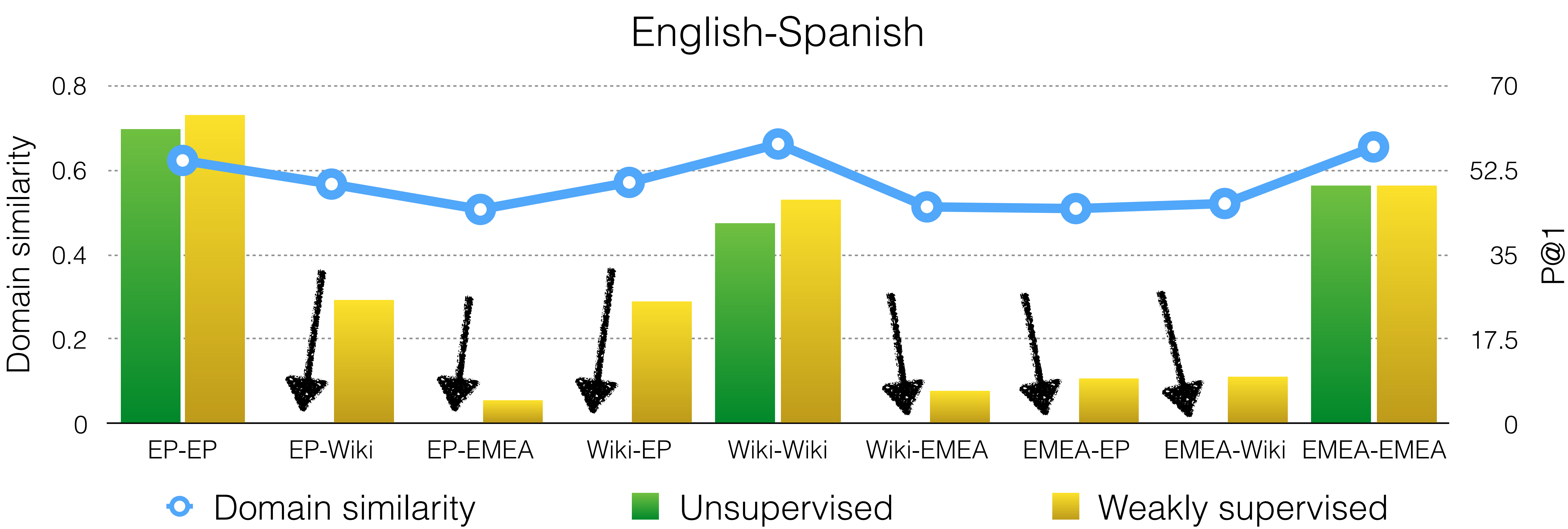
- Source and target embeddings induced on 3 corpora: EuroParl (EP), Wikipedia (Wiki), Medical (EMEA)



- Unsupervised approaches break down when domains are dissimilar

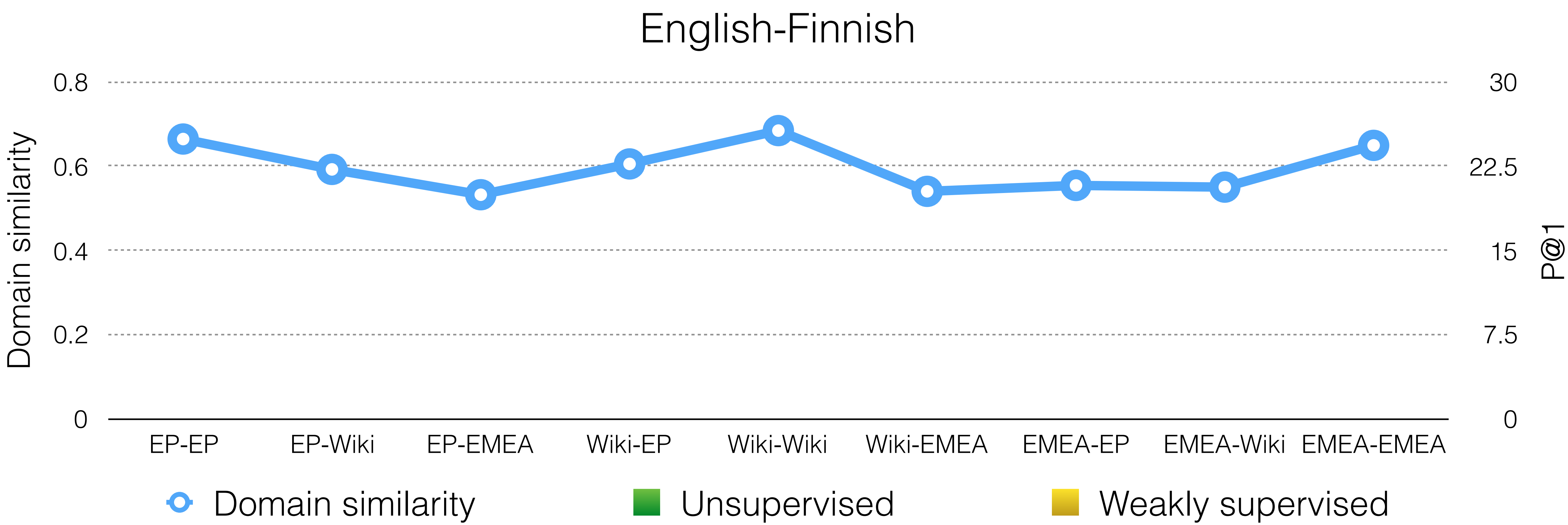
Impact of domain differences

- Source and target embeddings induced on 3 corpora: EuroParl (EP), Wikipedia (Wiki), Medical (EMEA)

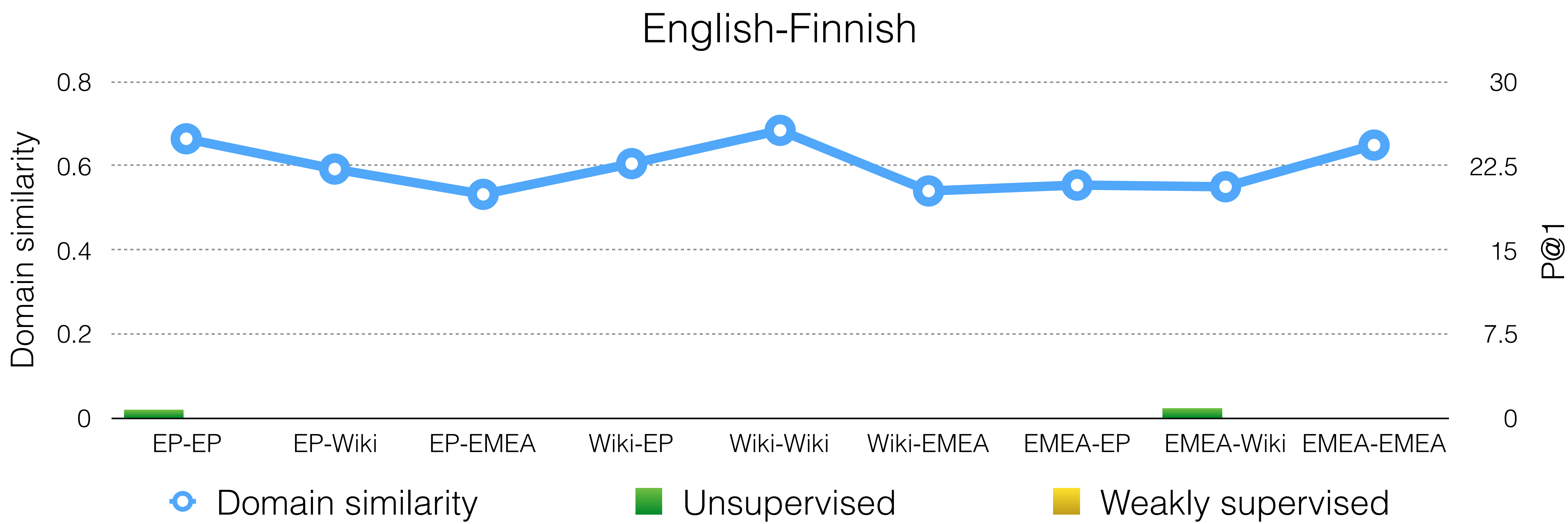


- Unsupervised approaches break down when domains are dissimilar

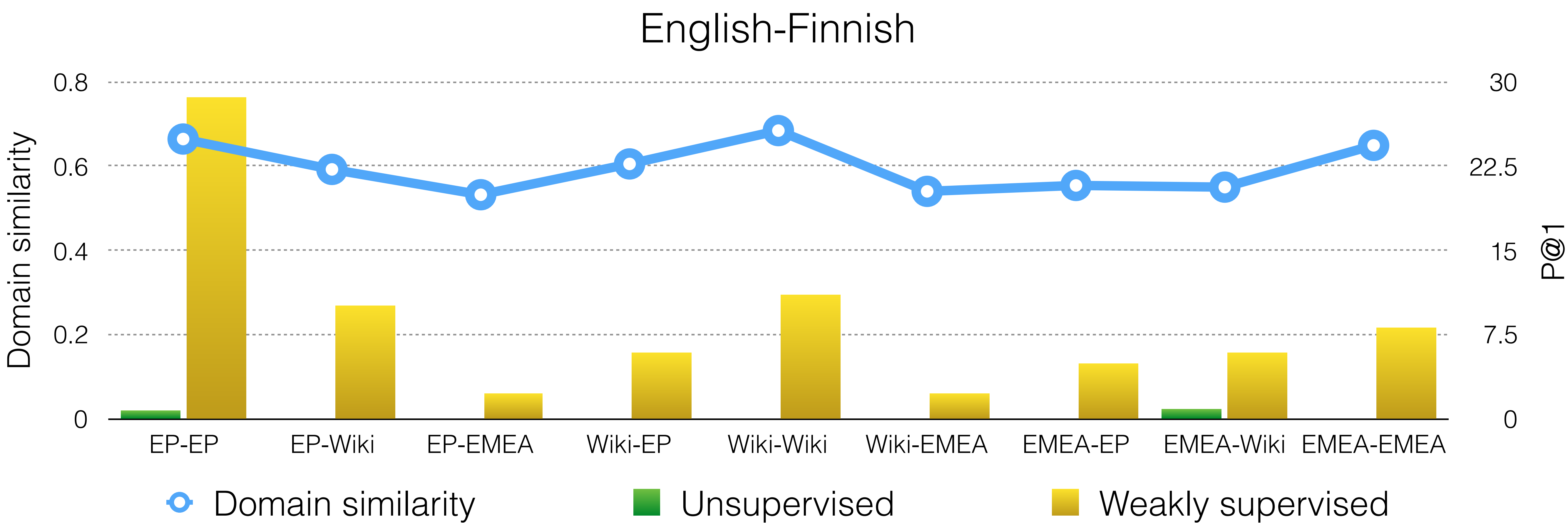
Impact of domain differences



Impact of domain differences

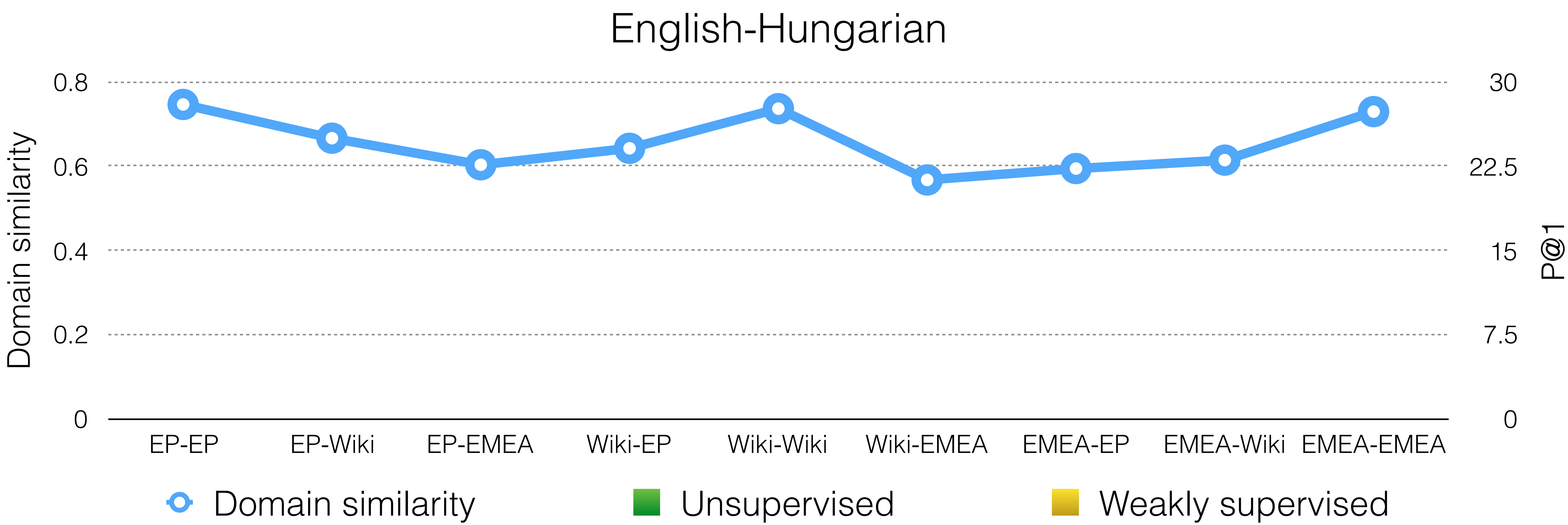


Impact of domain differences

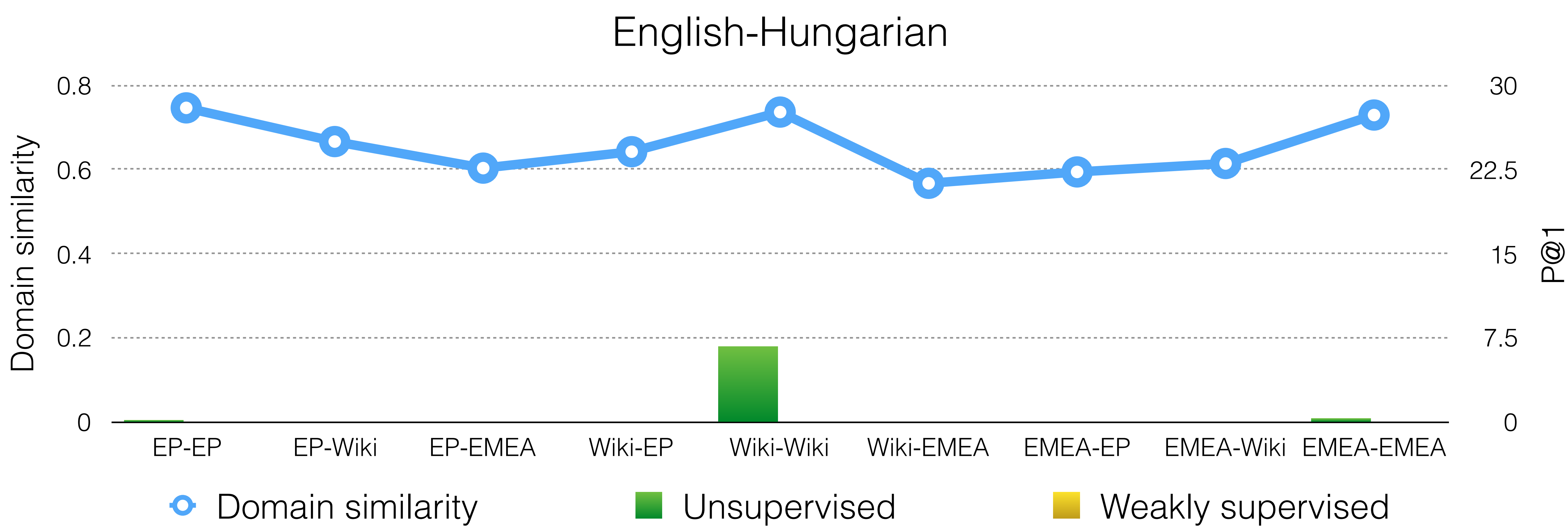


- Domain differences may exacerbate difficulties of generalising across dissimilar languages

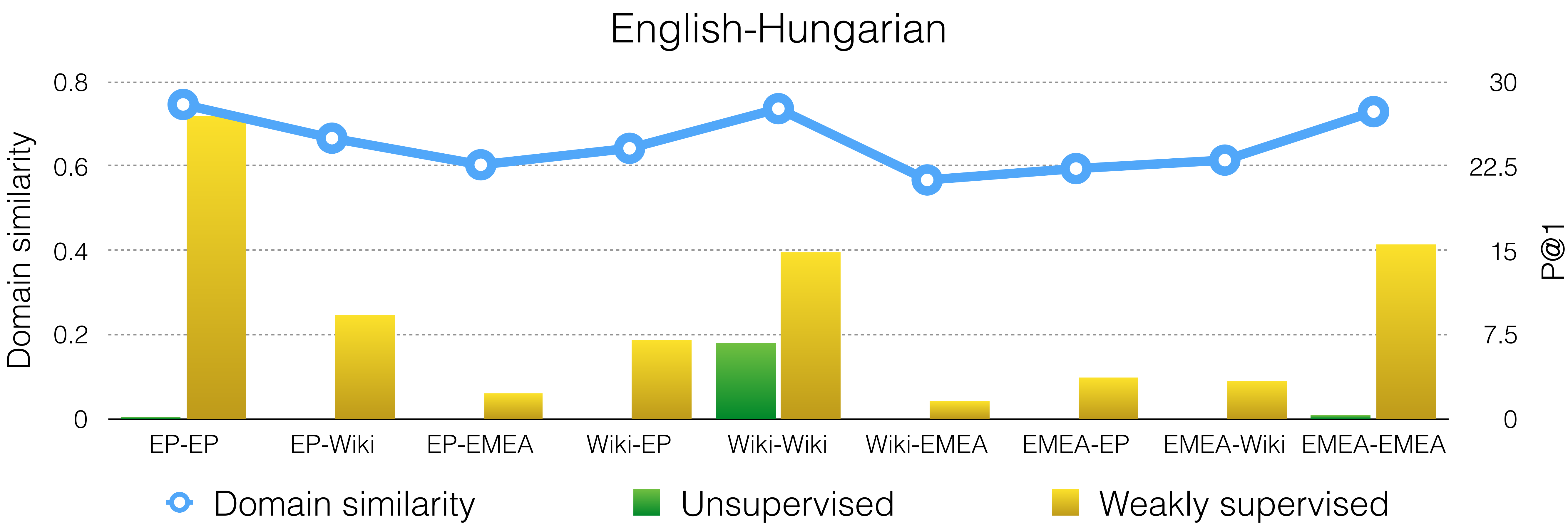
Impact of domain differences



Impact of domain differences



Impact of domain differences



- Weak supervision helps to bridge domain differences, but performance still deteriorates

Impact of hyper-parameters

Impact of hyper-parameters

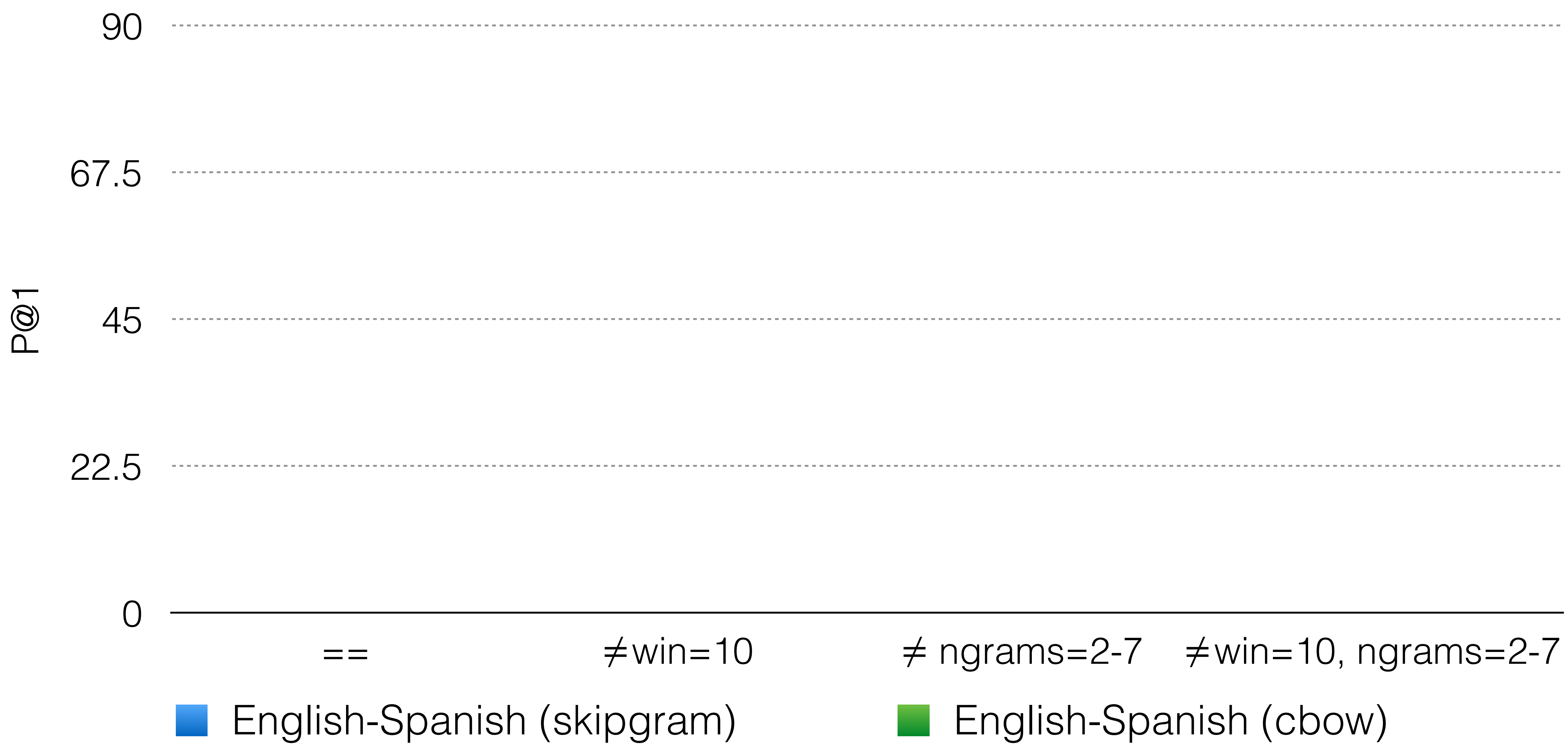
- Settings: English with skipgram, win=2, ngrams=3-6

Impact of hyper-parameters

- Settings: English with skipgram, win=2, ngrams=3-6
- Vary hyper-parameters of Spanish embeddings

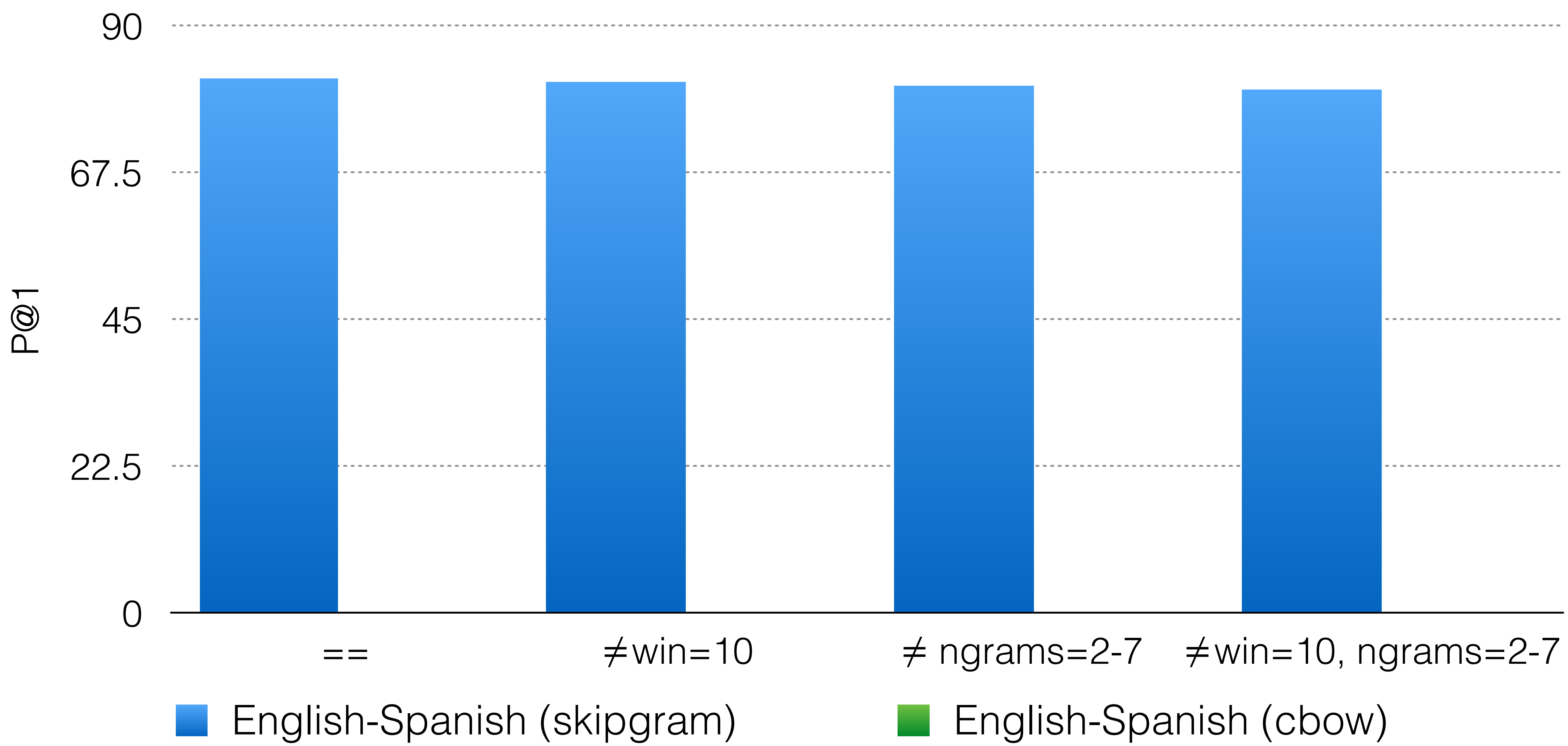
Impact of hyper-parameters

- Settings: English with skipgram, win=2, ngrams=3-6
- Vary hyper-parameters of Spanish embeddings



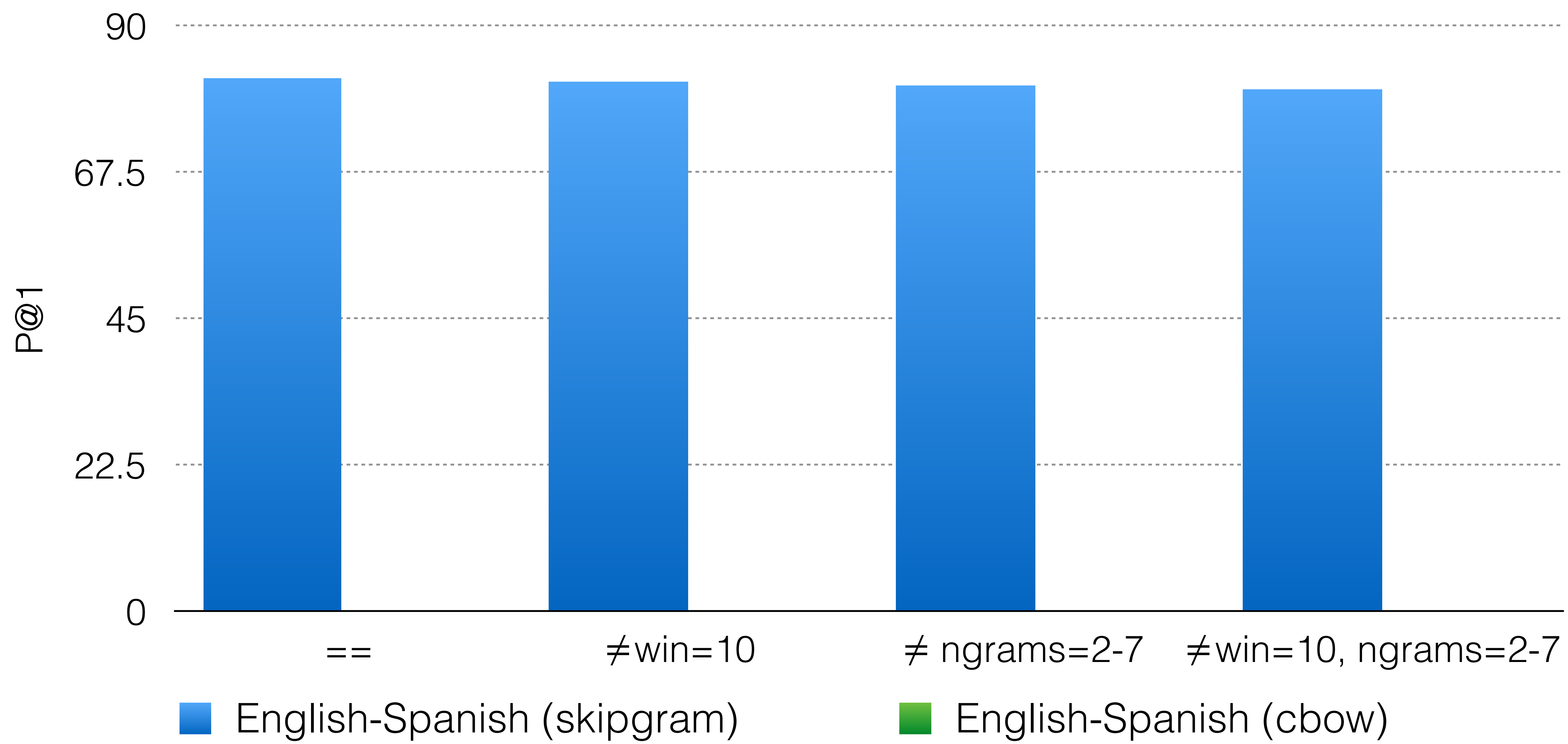
Impact of hyper-parameters

- Settings: English with skipgram, win=2, ngrams=3-6
- Vary hyper-parameters of Spanish embeddings



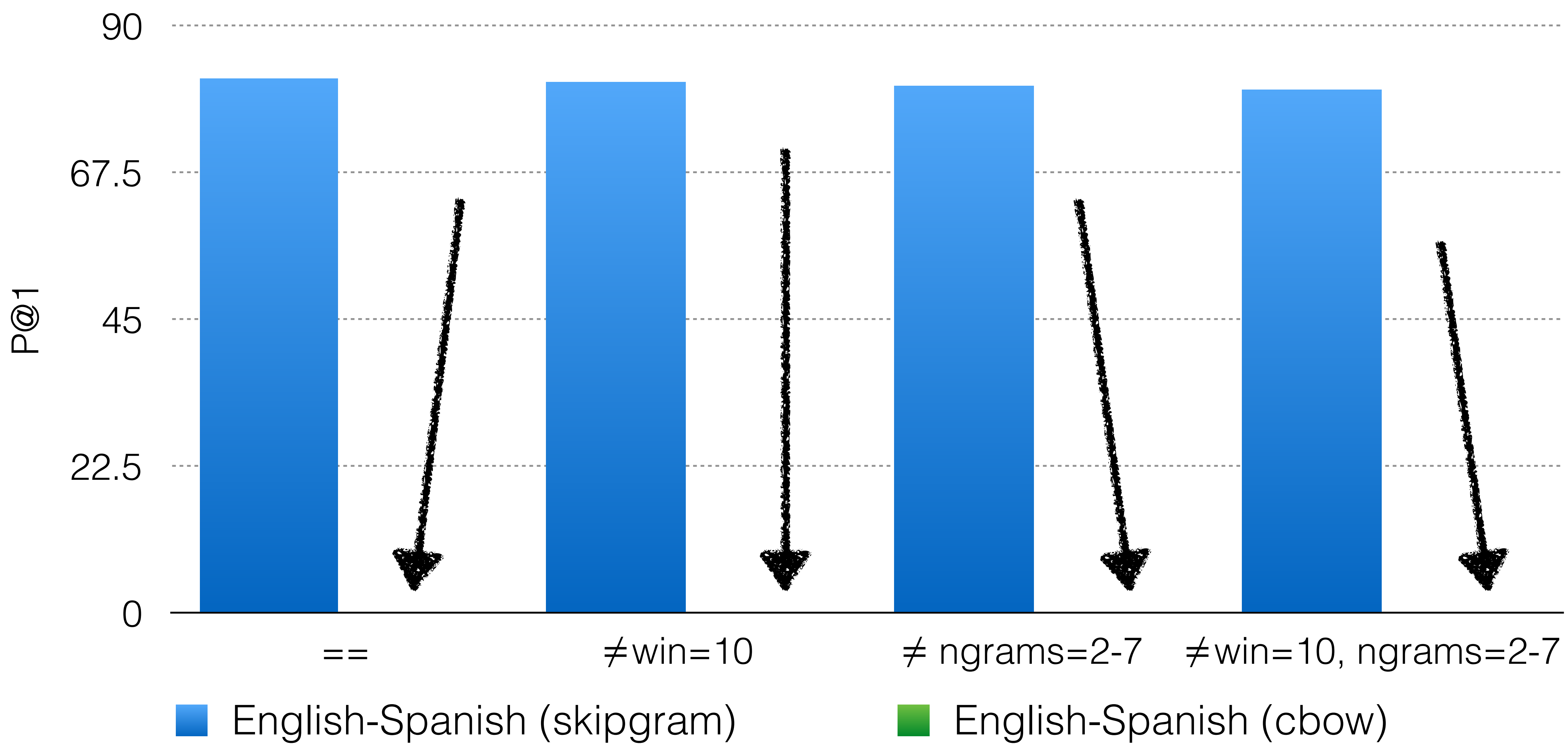
Impact of hyper-parameters

- Settings: English with skipgram, win=2, ngrams=3-6
- Vary hyper-parameters of Spanish embeddings

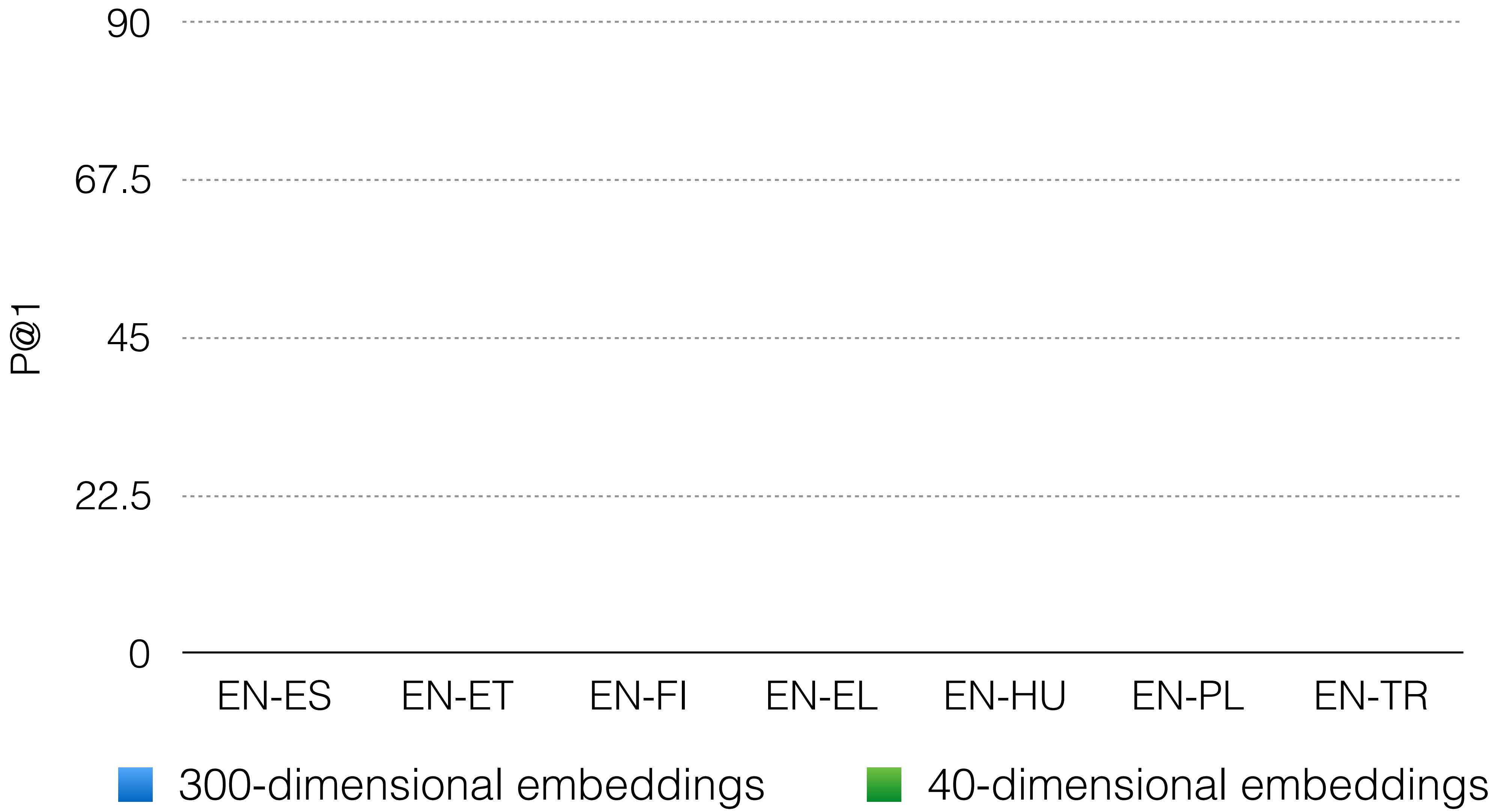


Impact of hyper-parameters

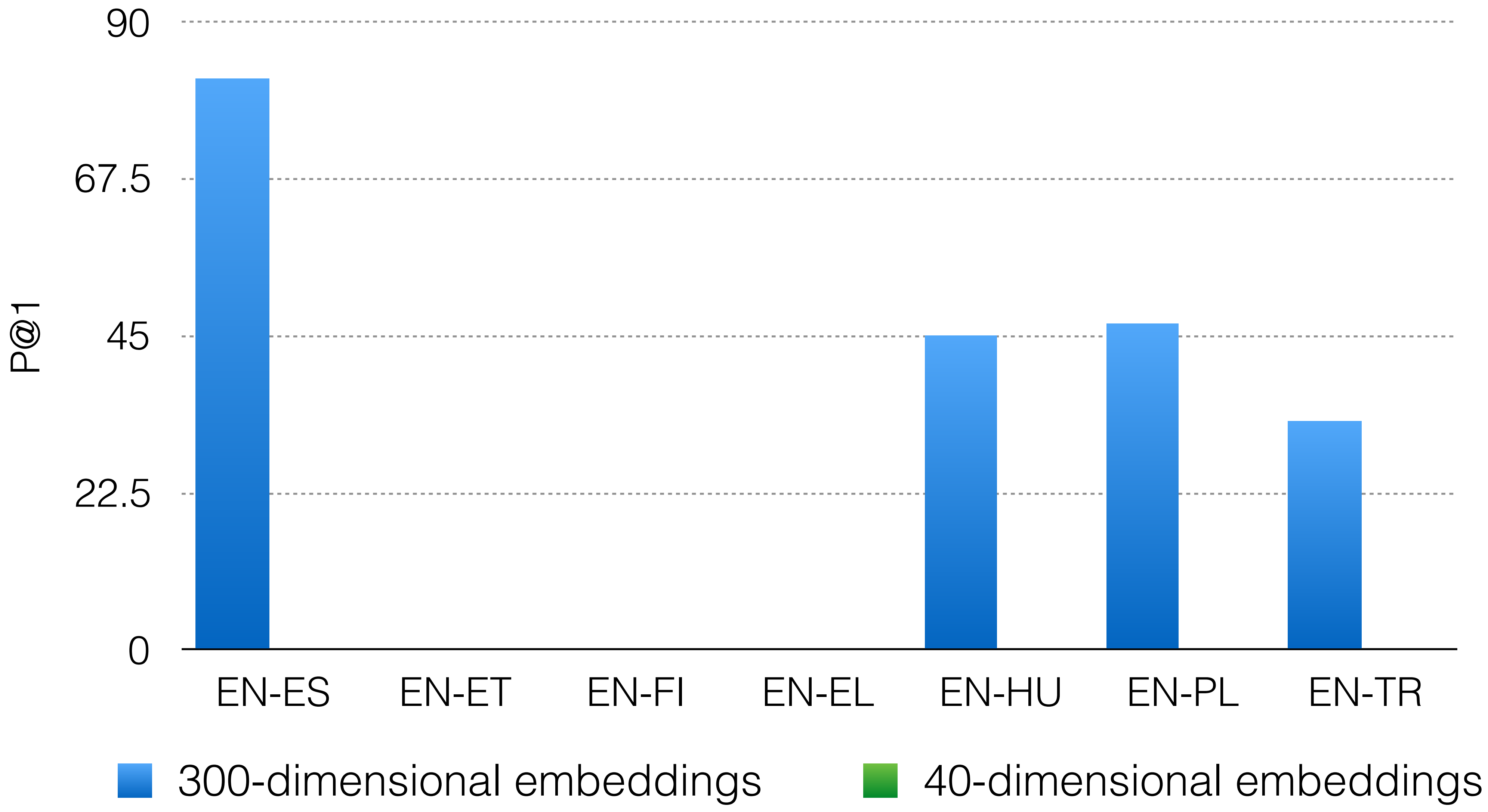
- **Different algorithms introduce embedding spaces with wildly different structures.**



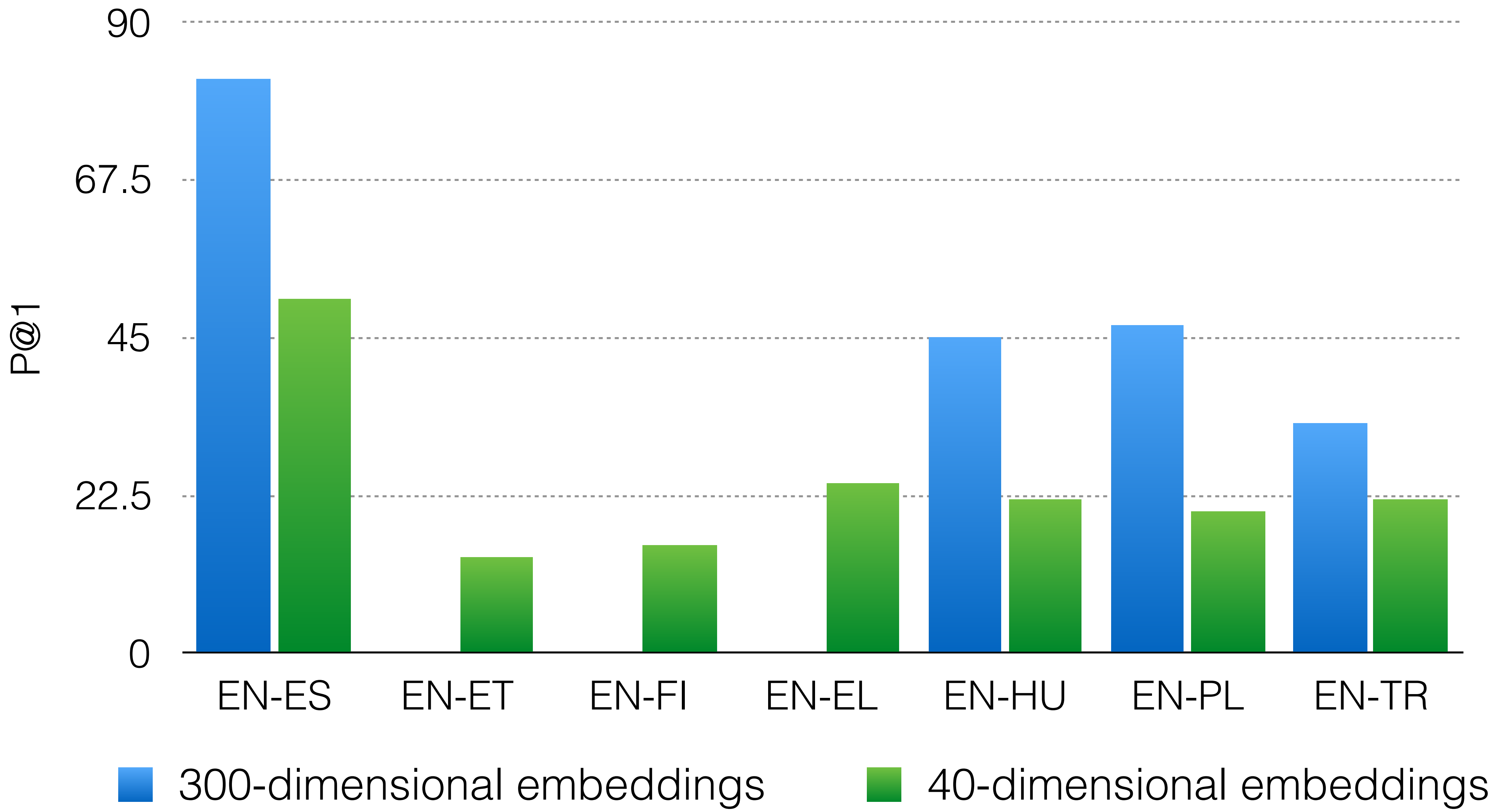
Impact of dimensionality



Impact of dimensionality

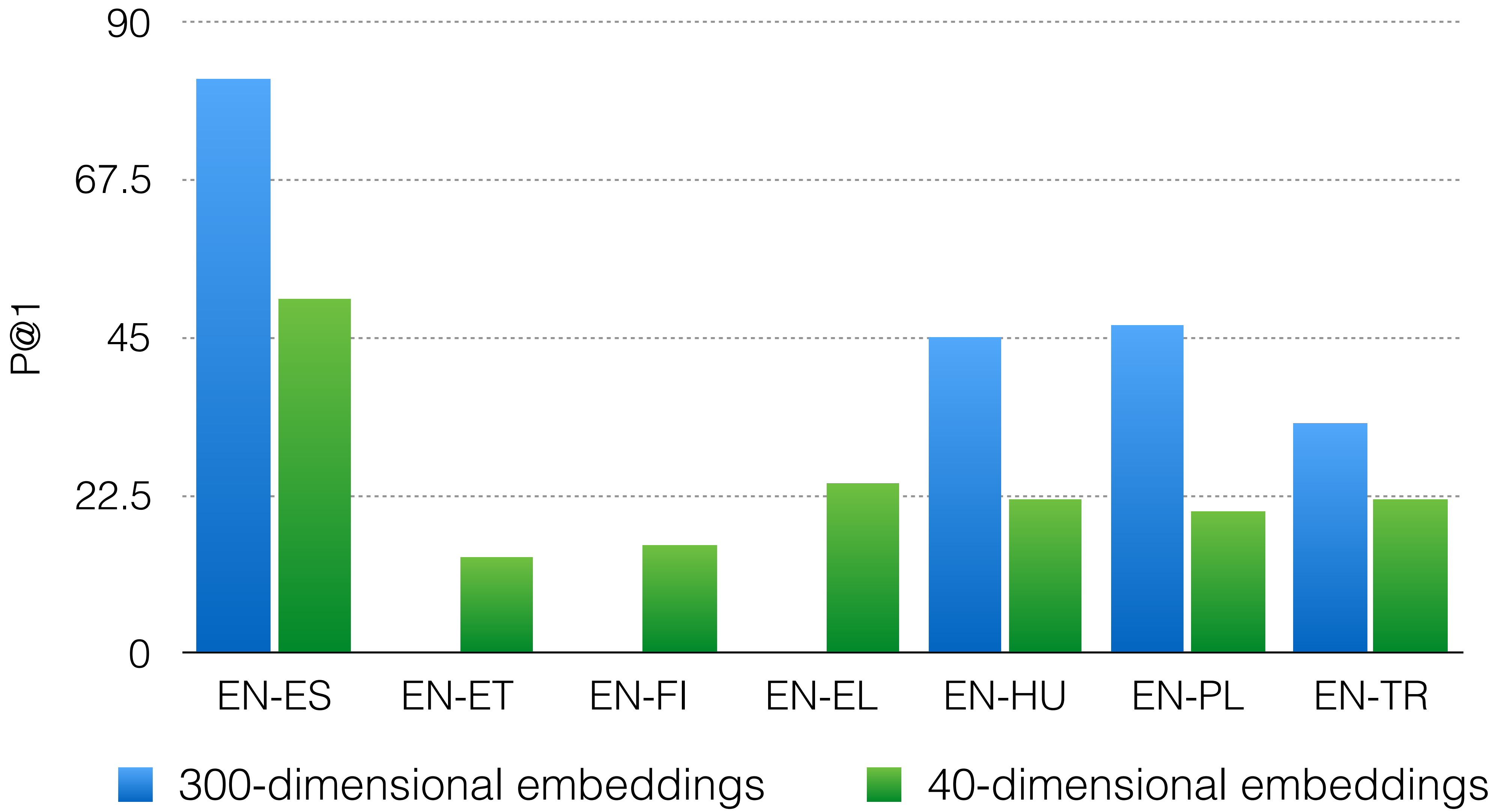


Impact of dimensionality



- Worse performance overall, but *better* performance for dissimilar language pairs (Estonian, Finnish, Greek).

Impact of dimensionality



- Worse performance overall, but *better* performance for dissimilar language pairs (Estonian, Finnish, Greek).
- **Monolingual word embeddings may overfit to rare peculiarities of languages.**

Impact of evaluation procedure

Impact of evaluation procedure

- **Part-of-speech:**
  Performance on verbs is lowest across the board.

Impact of evaluation procedure

- **Part-of-speech:**
  Performance on verbs is lowest across the board.
- **Frequency:**
  Sensitivity to frequency for Hungarian, but less so for Spanish.

Impact of evaluation procedure

- **Part-of-speech:**
  Performance on verbs is lowest across the board.
- **Frequency:**
  Sensitivity to frequency for Hungarian, but less so for Spanish.
- **Homographs:**
  Lower precision due to loan words/proper names. High precision for free with weak supervision.

Takeaways

Takeaways

- Word embedding spaces are **not approximately isomorphic** across languages.

Takeaways

- Word embedding spaces are **not approximately isomorphic** across languages.
- We can use **eigenvector similarity** to characterise the relatedness of two monolingual vector spaces.

Takeaways

- Word embedding spaces are **not approximately isomorphic** across languages.
- We can use **eigenvector similarity** to characterise the relatedness of two monolingual vector spaces.
- Eigenvector similarity **strongly correlates** with unsupervised bilingual dictionary induction performance.

Takeaways

- Word embedding spaces are **not approximately isomorphic** across languages.
- We can use **eigenvector similarity** to characterise the relatedness of two monolingual vector spaces.
- Eigenvector similarity **strongly correlates** with unsupervised bilingual dictionary induction performance.
- **Limitations** of unsupervised bilingual dictionary induction:

Takeaways

- Word embedding spaces are **not approximately isomorphic** across languages.
- We can use **eigenvector similarity** to characterise the relatedness of two monolingual vector spaces.
- Eigenvector similarity **strongly correlates** with unsupervised bilingual dictionary induction performance.
- **Limitations** of unsupervised bilingual dictionary induction:
  - **Morphologically rich** languages.

Takeaways

- Word embedding spaces are **not approximately isomorphic** across languages.
- We can use **eigenvector similarity** to characterise the relatedness of two monolingual vector spaces.
- Eigenvector similarity **strongly correlates** with unsupervised bilingual dictionary induction performance.
- **Limitations** of unsupervised bilingual dictionary induction:
  - **Morphologically rich** languages.
  - Corpora from **different domains**.

Takeaways

- Word embedding spaces are **not approximately isomorphic** across languages.
- We can use **eigenvector similarity** to characterise the relatedness of two monolingual vector spaces.
- Eigenvector similarity **strongly correlates** with unsupervised bilingual dictionary induction performance.
- **Limitations** of unsupervised bilingual dictionary induction:
  - **Morphologically rich** languages.
  - Corpora from **different domains**.
  - **Different word embedding** algorithms.