On the Role of Seed Lexicons in Learning Bilingual Word Embeddings

The work of the paper

- 1. Do **seed lexicon source** and **translation method** have influence on the performance?

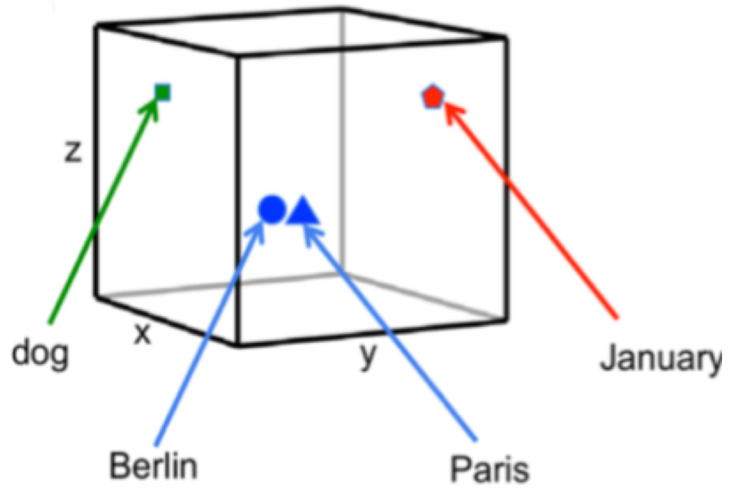
- 2. Can **seed lexicon size** influence the performance?

- 3. Does **translation pair reliability** really matter?

- 4. Based on the above analysis, this paper proposes a new *shared bilingual word embedding space* (SBWES) model only relying on **weak** or **inexpensive** document-level bilingual signals, but also can be used in **both** monolingual settings and bilingual settings.

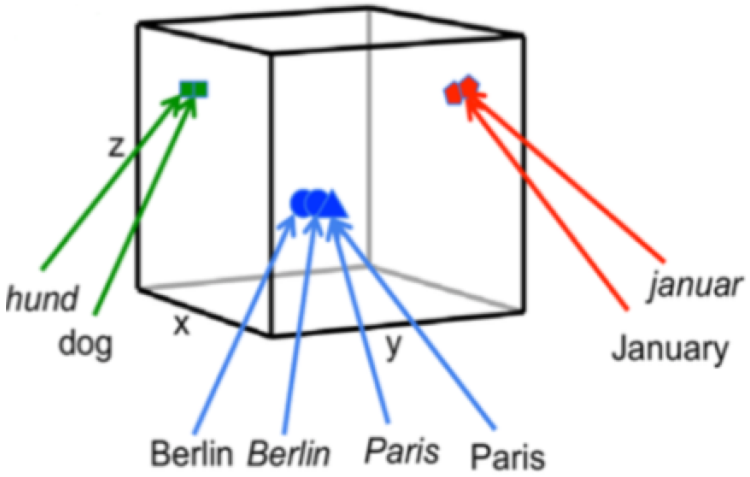
A common approach to shared bilingual word embedding space

$$\min_{\mathbf{W}\in\mathbb{R}^{d_S\times d_T}} ||\mathbf{XW}-\mathbf{Y}||_F^2+\lambda||\mathbf{W}||_F^2$$

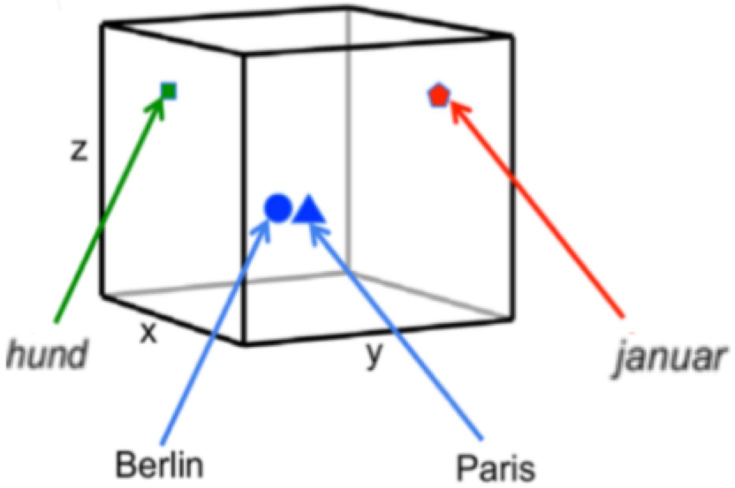


Monolingual

seed lexicons



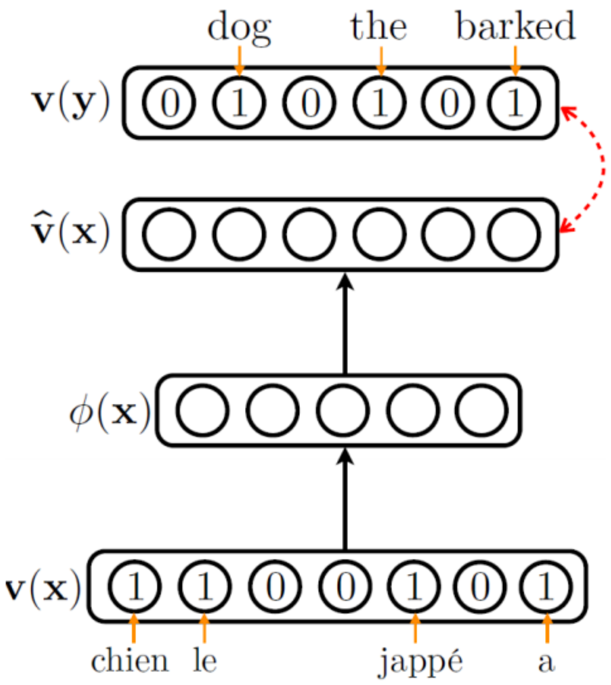
Bilingual

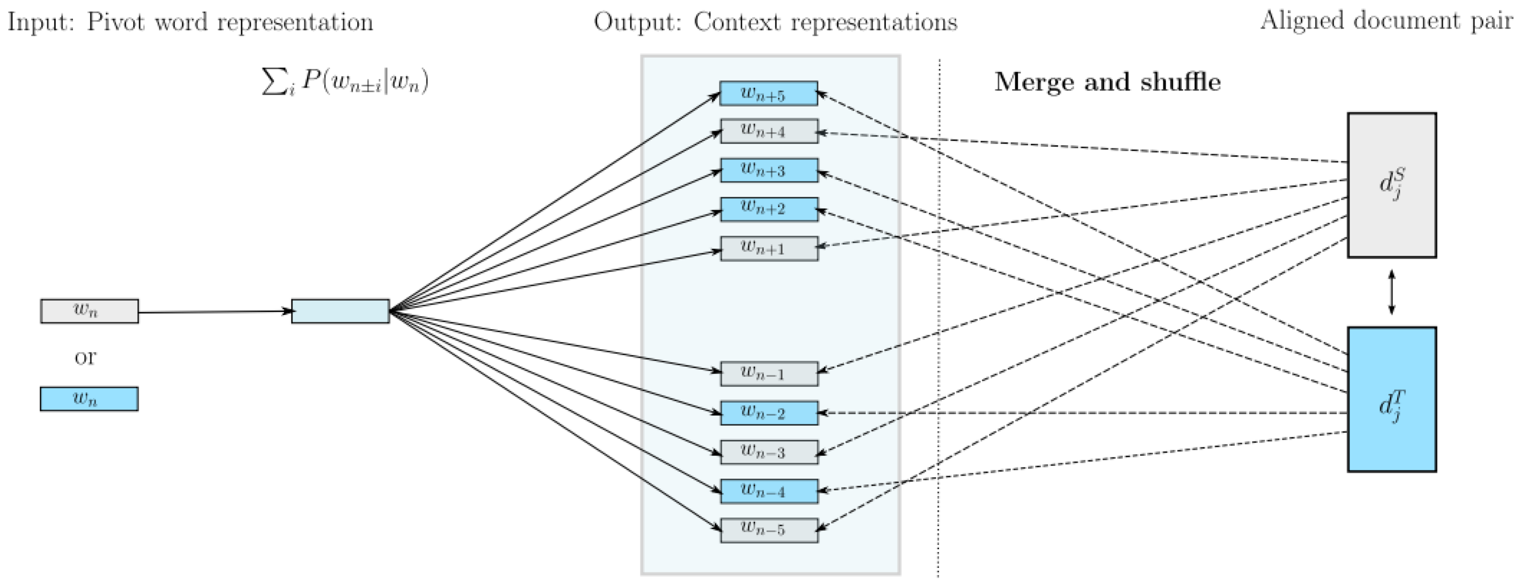


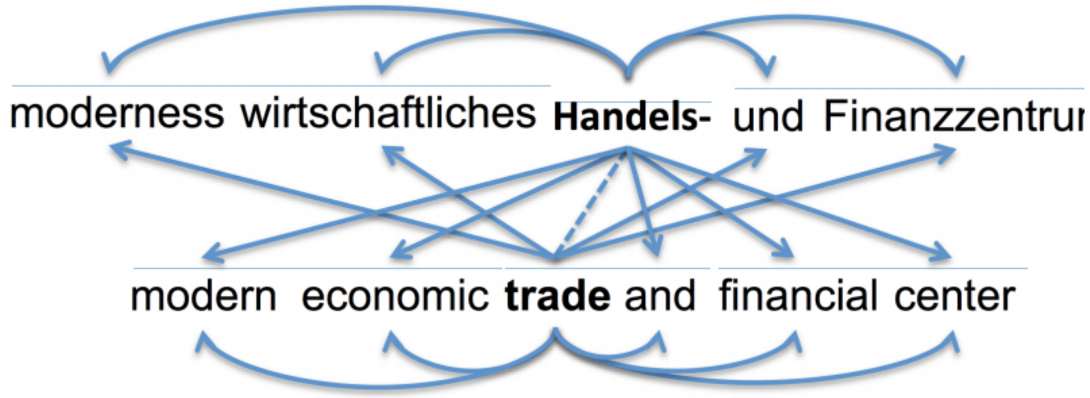
Monolingual

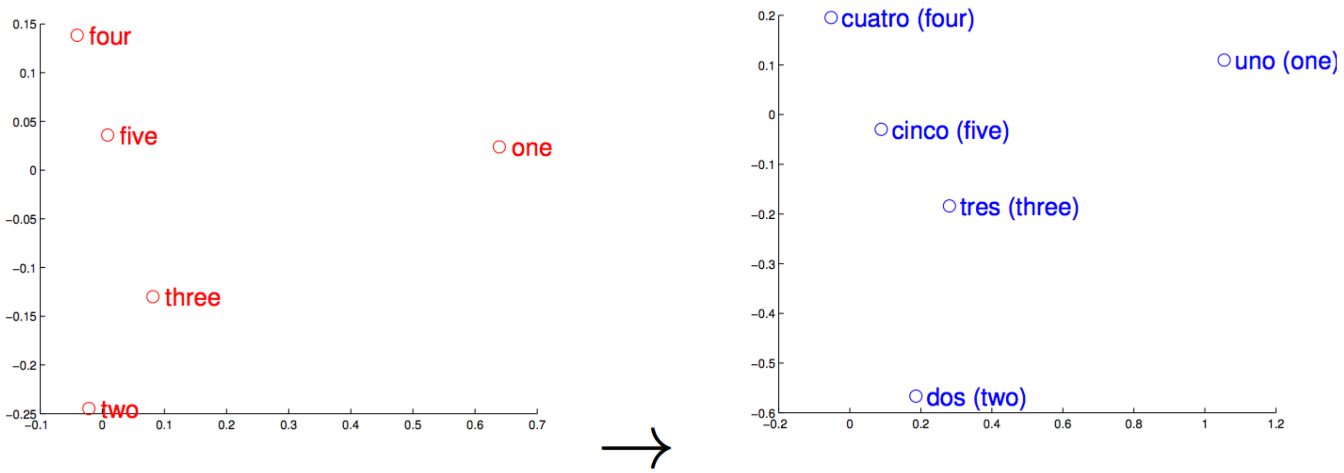
- Seed lexicon is a set containing word pairs. Every pair is a translation pair that involves word A, coming from one language, and A's corresponding translation word B from another language.

Previous BWE models






moderness wirtschaftliches Handels- und Finanzzentrur
modern economic trade and financial center



They either require expensive bilingual signals (word or sentence-level alignments) or cannot be simultaneously used in monolingual and bilingual settings.

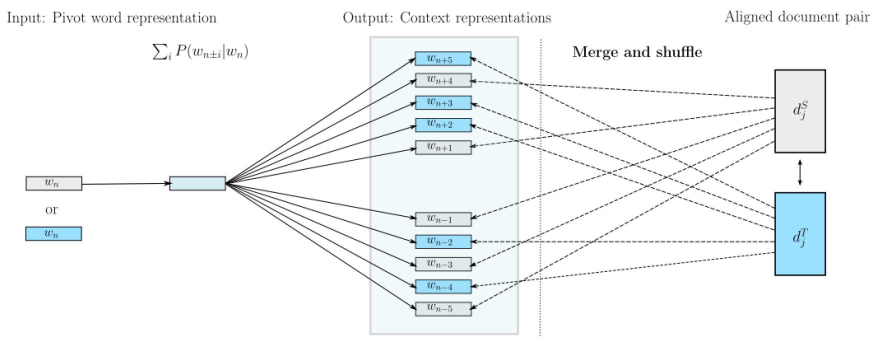
Hybrid model

- 1. Obtain SBWES-1 through Pseudo-Bilingual Document Training

- 2. Obtain seed lexicons through SBWES-1

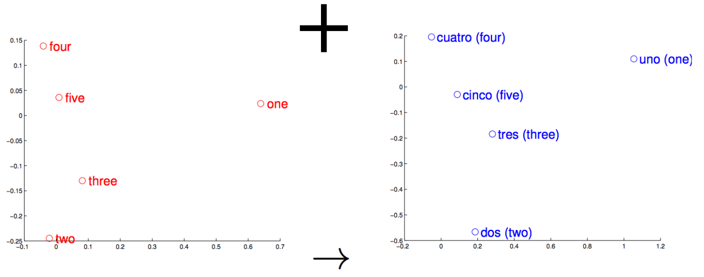
- 3. Obtain SBWES-2, namely, the final SBWES, through Post-Hoc Mapping with seed lexicons

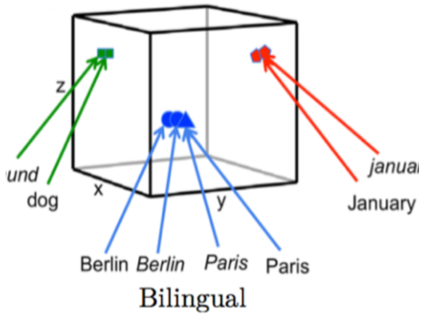
$$\min_{\mathbf{W}\in\mathbb{R}^{d_S\times d_T}} ||\mathbf{XW}-\mathbf{Y}||_F^2 + \lambda||\mathbf{W}||_F^2$$



seed lexicon

+



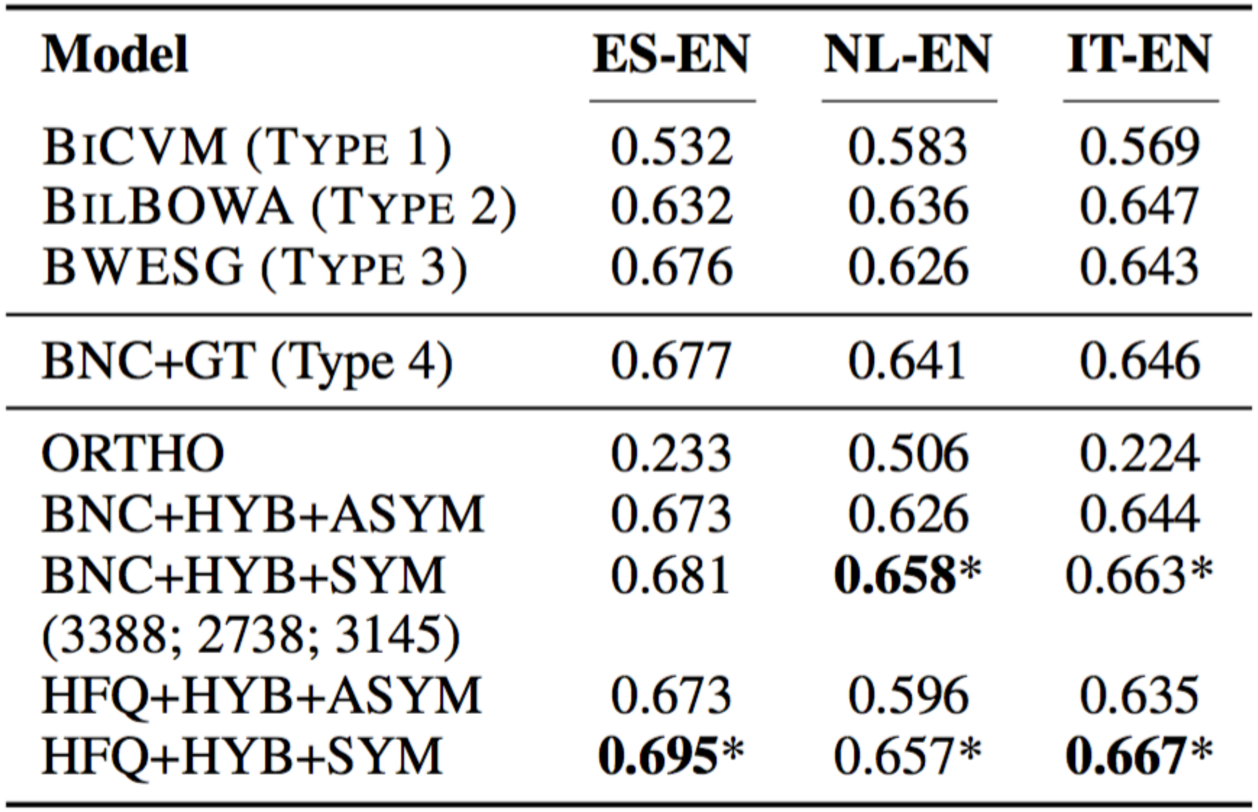


Some abbreviations

- Seed Lexicon Source:
  - **BNC** — a list containing 6,318 most frequent English lemmas
  - **HFQ** — the TOP-N most frequent words occurring in training corpora
  - **ORTHO** — all words shared between two monolingual vocabularies
- Translation Method:
  - **GT** — Google Translate
  - **HYB** — translation by SBWES-1 using the nearest neighbour distance
- Translation Pair Reliability:
  - **SYM** — symmetry constraint: two words are used as seed lexicon pairs only if they are mutual nearest neighbours given their representations in SBWES-1
  - **ASYM** — not adding symmetry constraint

Standard bilingual lexicon learning task (I)

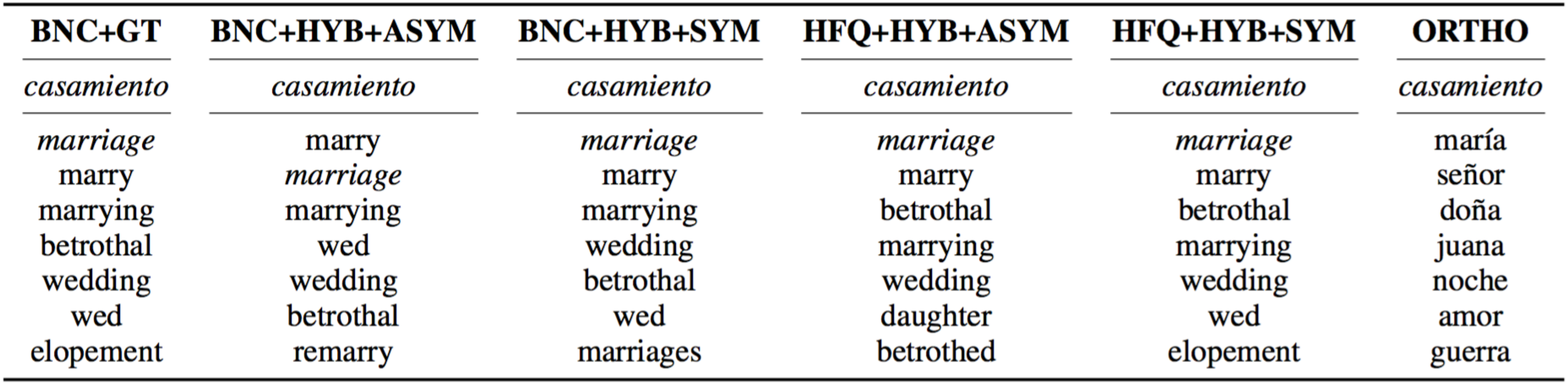
- Evaluation Metrics: We measure the BLL performanc using the standard *Top 1 accuracy*.
- Careful selection of reliable pairs can lead to peak performances even with a lower number of pairs.



Model	**ES-EN**	**NL-EN**	**IT-EN**
BiCVM (Type 1)	0.532	0.583	0.569
BiLBOWA (Type 2)	0.632	0.636	0.647
BWESG (Type 3)	0.676	0.626	0.643
BNC+GT (Type 4)	0.677	0.641	0.646
ORTHO	0.233	0.506	0.224
BNC+HYB+ASYM	0.673	0.626	0.644
BNC+HYB+SYM (3388; 2738; 3145)	0.681	**0.658***	0.663*
HFQ+HYB+ASYM	0.673	0.596	0.635
HFQ+HYB+SYM	**0.695***	0.657*	**0.667***

Standard bilingual lexicon learning task (II)

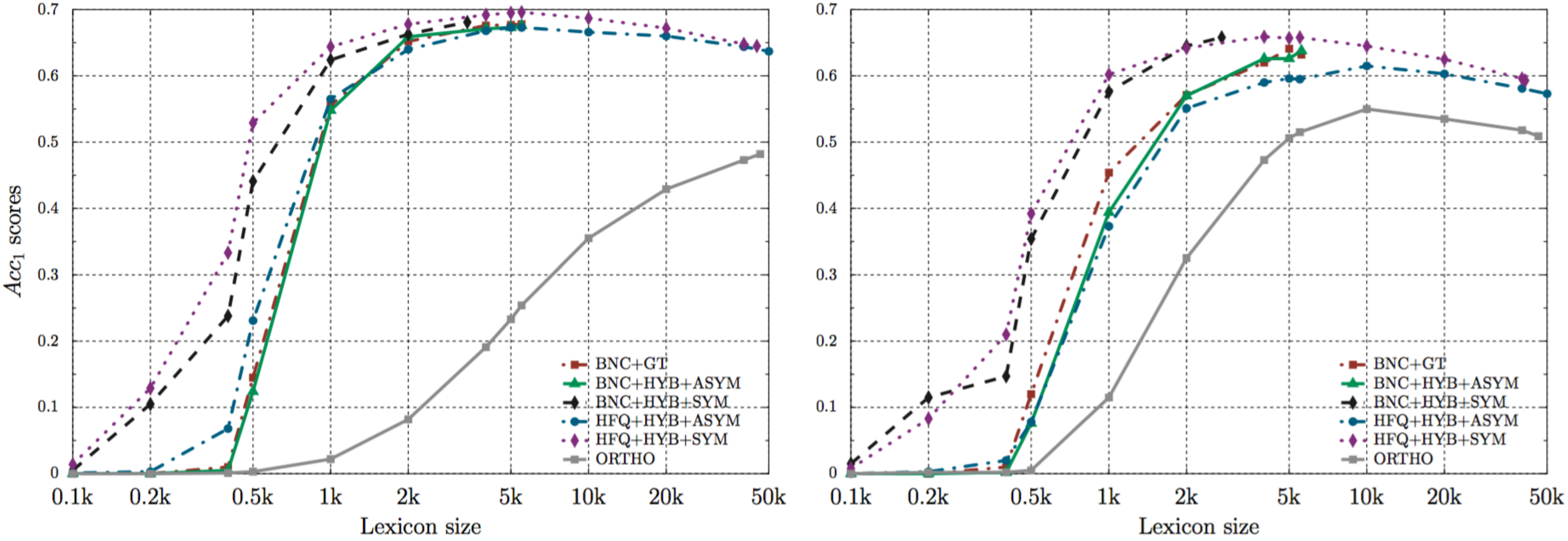
- The choice of seed lexicon pairs may strongly influence the properties of the SBWES.



BNC+GT	**BNC+HYB+ASYM**	**BNC+HYB+SYM**	**HFQ+HYB+ASYM**	**HFQ+HYB+SYM**	**ORTHO**
casamiento	*casamiento*	*casamiento*	*casamiento*	*casamiento*	*casamiento*
marriage	marry	*marriage*	*marriage*	*marriage*	maría
marry	*marriage*	marry	marry	marry	señor
marrying	marrying	marrying	betrothal	betrothal	doña
betrothal	wed	wedding	marrying	marrying	juana
wedding	wedding	betrothal	wedding	wedding	noche
wed	betrothal	wed	daughter	wed	amor
elopement	remarry	marriages	betrothed	elopement	guerra

Lexicon size

- Do not blindly use all potential training pairs, but rely on the reliable ones.
- Google translate can be safely replaced by a document-level embedding model.



(a) Spanish-English

(b) Dutch-English

Reference

- [1] Kiros R, Zhu Y, Salakhutdinov R R, et al. Skip-thought vectors[C]//Advances in neural information processing systems. 2015: 3294-3302.

- [2]Mikolov T, Le Q V, Sutskever I. Exploiting similarities among languages for machine translation[J]. arXiv preprint arXiv:1309.4168, 2013.

- [3] Vulic I, Moens M F. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015). ACL, 2015.

- [4] Lauly S, Larochelle H, Khapra M, et al. An autoencoder approach to learning bilingual word representations[C]//Advances in Neural Information Processing Systems. 2014: 1853-1861.

- [5] Luong T, Pham H, Manning C D. Bilingual word representations with monolingual quality in mind[C]//Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. 2015: 151-159.