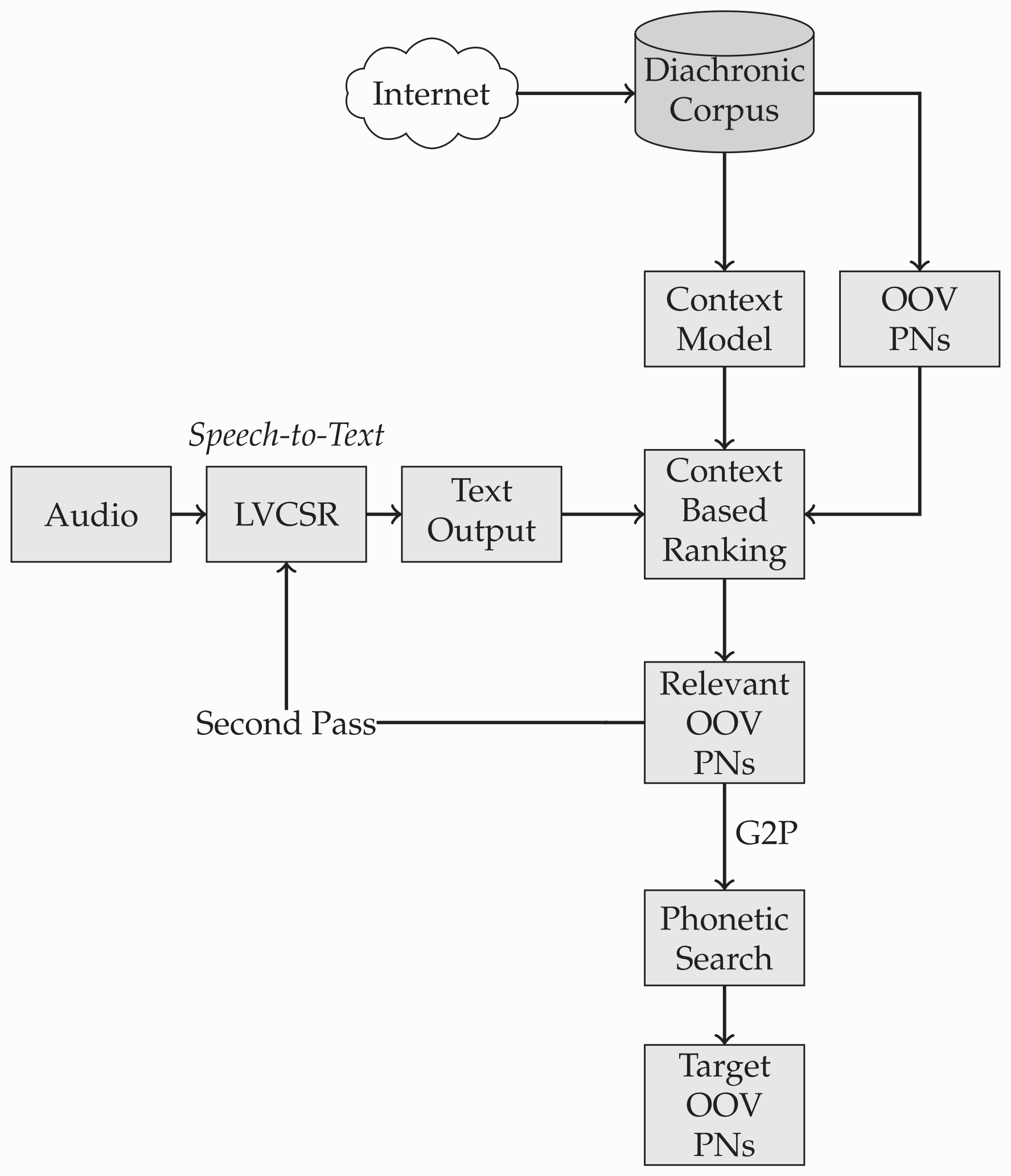
Document Level Semantic Context For Retrieving OOV Proper Names

Imran Sheikh*+, Irina Illina*, Dominique Fohr*, Georges Linarès+
*Multispeech, LORIA-INRIA, 54500 Villers-lès-Nancy, France
+LIA, University of Avignon, 84911 Avignon, France

Background

- *Out-Of-Vocabulary* (OOV) words in *Large Vocabulary Continuous Speech Recognition* (LVCSR)
- Most OOV in audio news are *Proper Names* (PNs)
- (Topic) context can refine recovery of PNs (Sheikh et. al., in *IEEE ICASSP*, 2015)



- Focus: retrieving semantically relevant OOV PNs

PN Context Vector Approach

- Training Phase
  - Learn semantic vector representation of each OOV PN from diachronic corpus
  - E.g. LDA based topic distribution vectors (Sheikh et. al., in *IEEE ICASSP*, 2015)
- Testing Phase
  - Get LVCSR hypothesis of audio document
  - Infer semantic vector representation of LVCSR hypothesis (h)
  - Compare LVCSR hypothesis and OOV PNs in semantic vector space
  - For LDA topic space

$$p(oov|h) = \sum_{t=1}^{T} p(oov|t)\; p(t|h)$$

Problem: Co-occurrence based topic/semantic context models biased against less frequent OOV PNs

Proposed Solution: Document specific context vectors for OOV PNs

Document Context Approach

- Training Phase
  - Learn vector representation of each document (d) in diachronic corpus
  - Store document vector as *one of the context vectors* of OOV PN in that document (d)
- Testing Phase
  - Infer vector representation of the LVCSR word hypothesis (h)
  - Compare context vector of LVCSR hypothesis with those for each OOV PN
  - Scoring function for jth OOV PN with C context vectors:

$$\max_{i \in C}\{CosineSimilarity(h, c_i^j)\}$$

Works for different document representations:

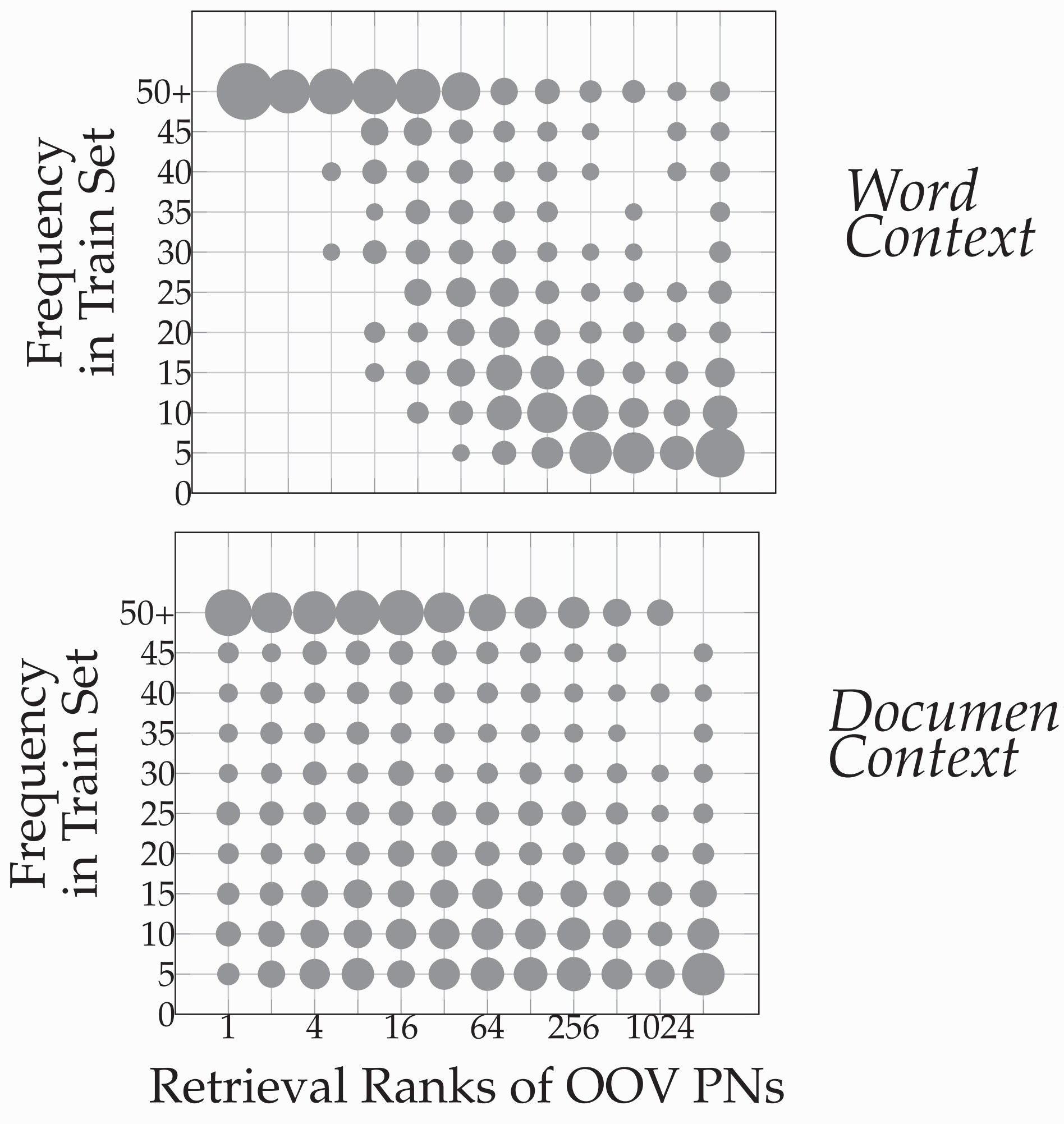
- Random Projection (RP) of tf-idf vectors
- LDA and LSA documents vectors
- Average of neural word vectors, CBOW & Skip-gram (Mikolov 2013), GloVe (Pennington 2014)

Experiments and Results

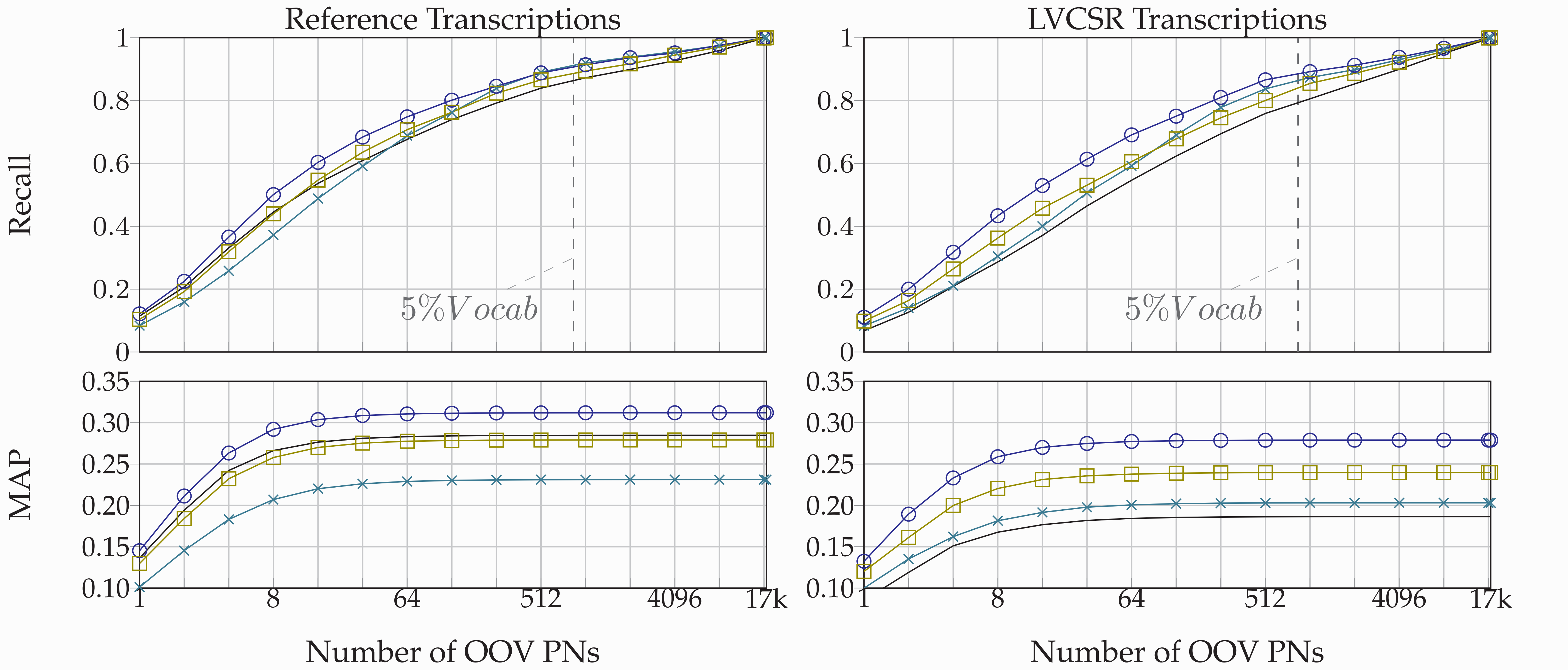
- Diachronic (L'Express) and Test (Euronews) corpus

	L'Express	Euronews
Type of Documents	Text	Video
Time Period	Jan 2014 - Jun 2014	
Number of Documents	45K	3K
Corpus Size (total word count)	24M	600K
Number of PN unigrams+	40K	2.2K
Number of OOV PN unigrams+	17K	1024
Documents with OOV PN	36K	1415

- Rank-Frequency Distribution of OOV PN with LDA



- OOV PN retrieval performance (—— RP, —×— LDA, —○— CBOW, —□— GloVe); LSA~LDA, Skip-gram~CBOW



Conclusion

- Document level semantic representations improve retrieval of less frequent OOV PNs.
- Retrieval performance trend for different representations: CBOW/Skip-gram > GloVe > LDA/LSA > RP
- A phonetic search for target OOV PNs confirmed that the retrieval is reliable for recovery of OOV PNs

This work is funded by the ContNomina project supported by the French National Research Agency (ANR) under contract ANR-12-BS02-0009.