SENSEMBERT: CONTEXT-ENHANCED SENSE EMBEDDINGS FOR MULTILINGUAL WORD SENSE DISAMBIGUATION

BIANCA SCARLINI, TOMMASO PASINI AND ROBERTO NAVIGLI
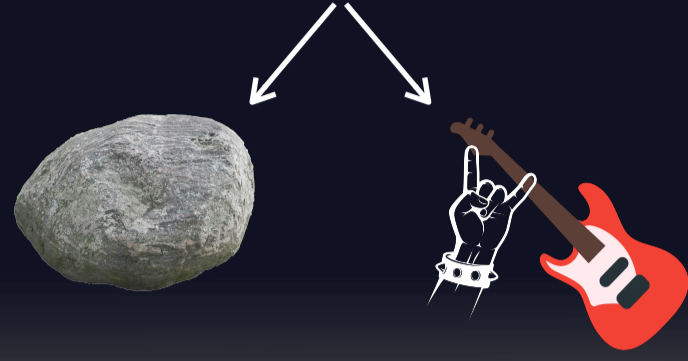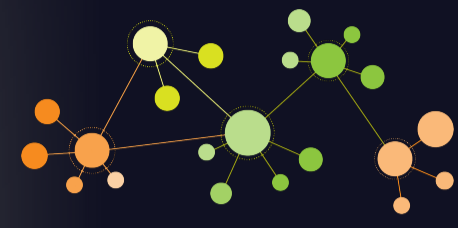
http://sensembert.org

WORD SENSE DISAMBIGUATION

This guitar is good for playing **rock**.

CONTRIBUTIONS

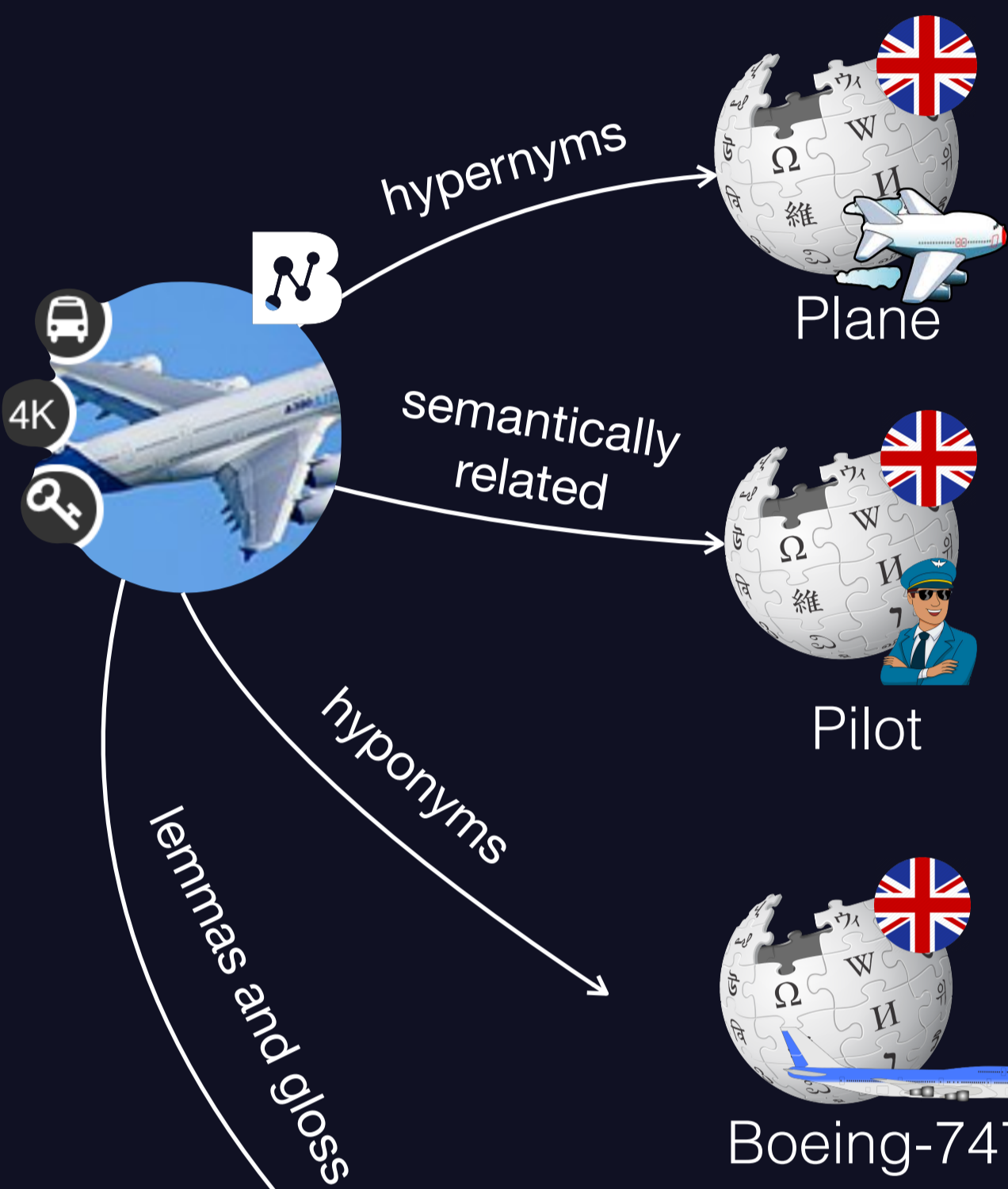**Knowledge-based** approach for building sense embeddings in **multiple languages**.

The embeddings take into account the **contexts** where a sense may appear.

The embeddings lay in a latent space that is **comparable** with the one of **BERT** contextual word embeddings.

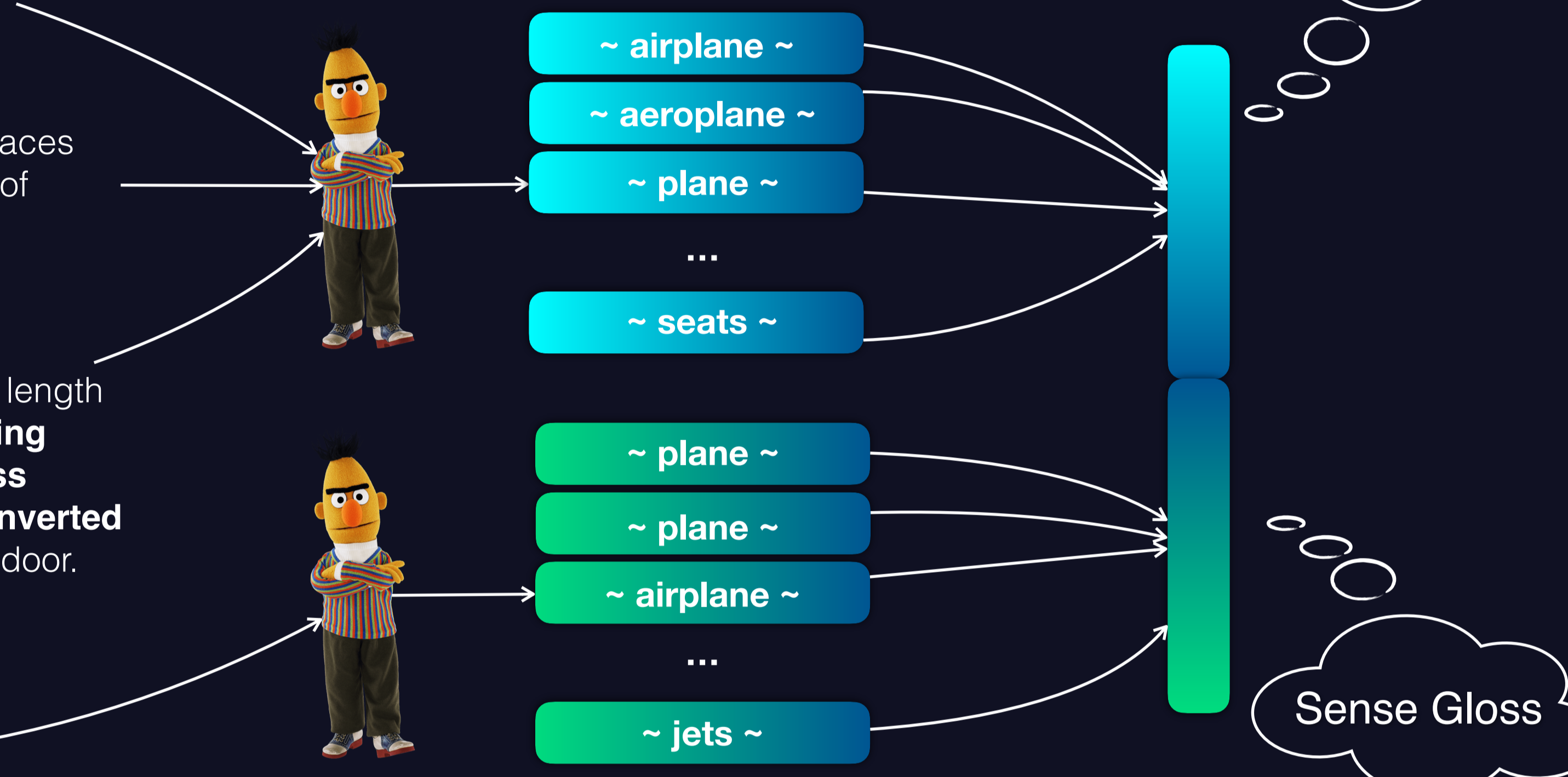## METHOD

Relevant words

hypernyms

An **airplane** or **aeroplane** (informally **plane**) is a powered, **fixed-wing aircraft** that is **propelled** forward by thrust from a **jet engine**, propeller or **rocket engine**.

Plane

semantically related

The number of **airline pilots** could decrease as **automation** replaces **copilots** and eventually **pilots** as well. […] the number and kind of **flying jobs** available may be decreased.

Pilot

hyponyms

The **quadjet 747** uses a **double-deck** configuration for part of its length and is available in **passenger**, freighter, and other versions. **Boeing** designed the **747**'s hump-like upper **deck** to serve as a **first-class lounge** or extra **seating**, and to allow the **aircraft** to be easily **converted** to a **cargo** carrier by removing **seats** and installing a front cargo door.
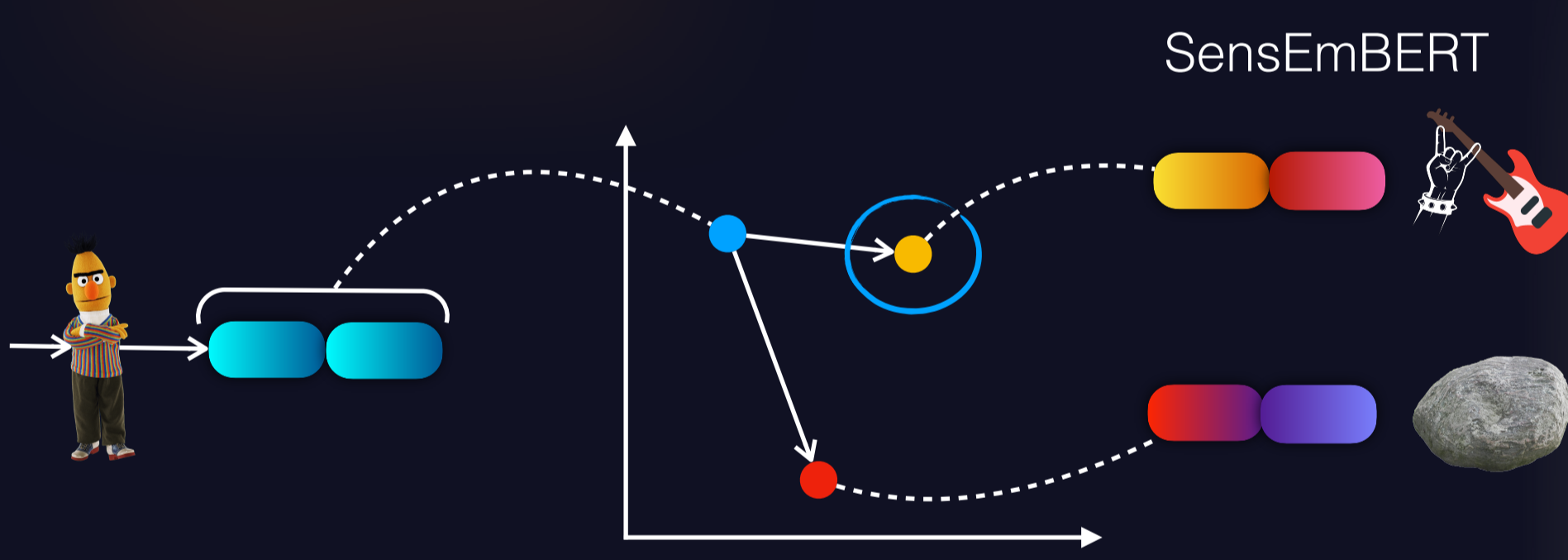
Boeing-747

lemmas and gloss

plane - plane, airplane. An aircraft that has a fixed wing and is powered by propellers or jets.

~ airplane ~
~ aeroplane ~
~ plane ~
...
~ seats ~

~ plane ~
~ plane ~
~ airplane ~
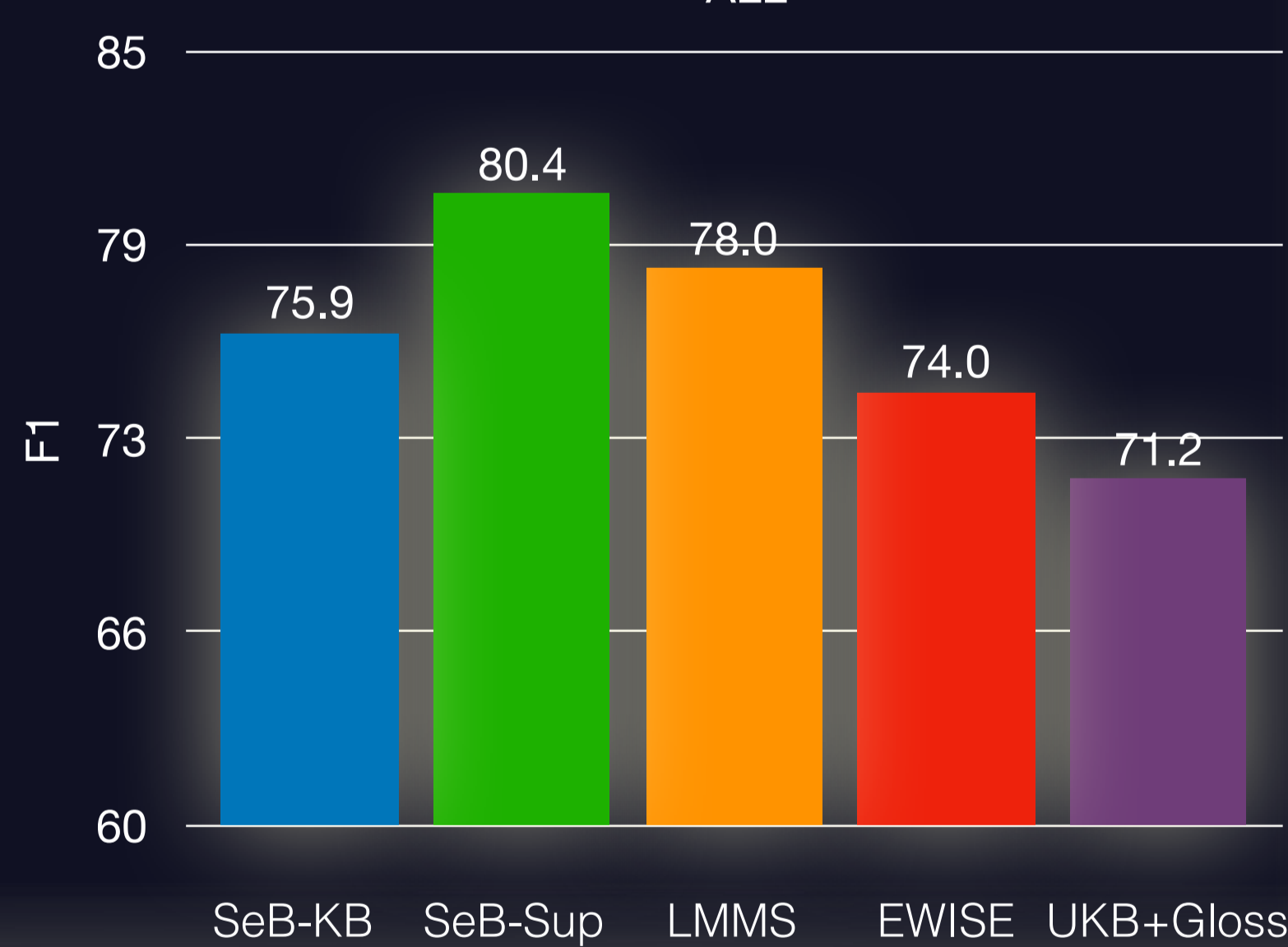...
~ jets ~

Context

Sense Gloss

ENGLISH WSD

This guitar is good for playing **rock**.

SensEmBERT

Competing Systems:

SeB-KB: SensEmBERT embeddings.
SeB-Sup: Combines SemCor contexts with SeB-KB embeddings.
LMMS: Sense embeddings learned from SemCor.
EWISE: Neural supervised approach.
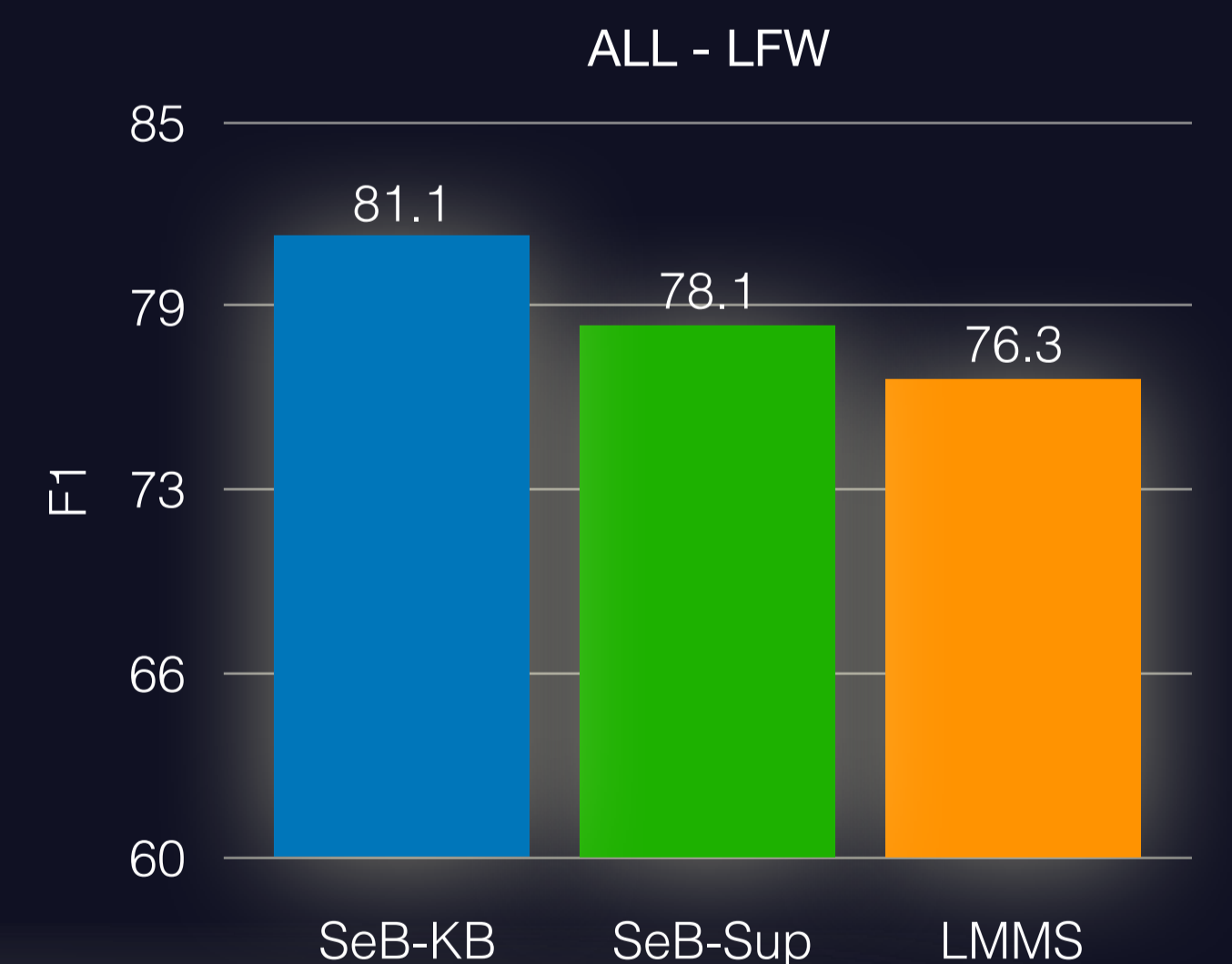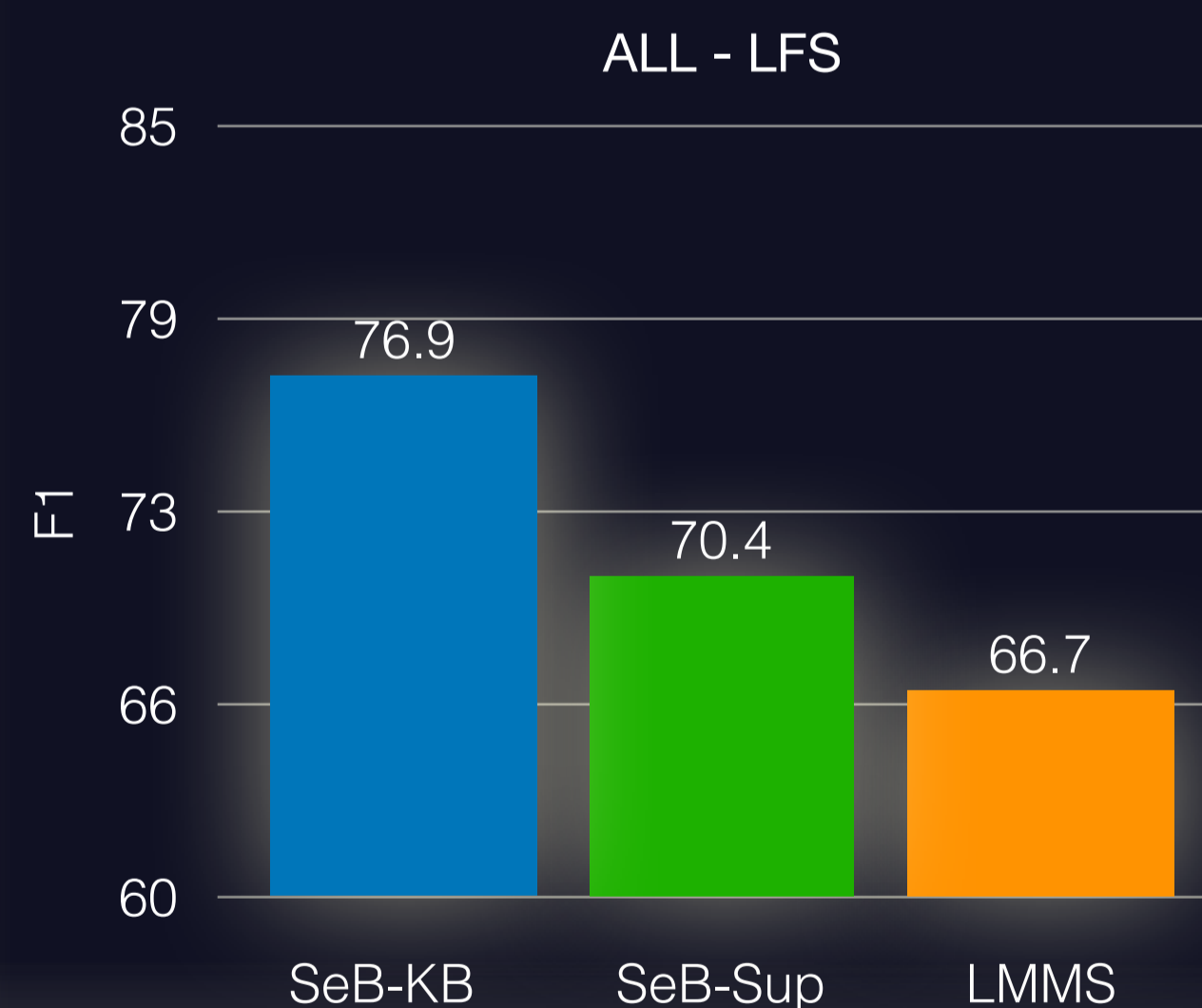**UKB+Gloss**: A knowledge-based algorithm exploiting glosses and semantic relations.

ALL

System	F1
SeB-KB	75.9
SeB-Sup	80.4
LMMS	78.0
EWISE	74.0
UKB+Gloss	71.2

RARE INSTANCES

ALL - LFS (Least Frequent Senses):
Dataset containing all the instances in ALL tagged with a sense not in SemCor.

ALL - LFW (Least Frequent Words):
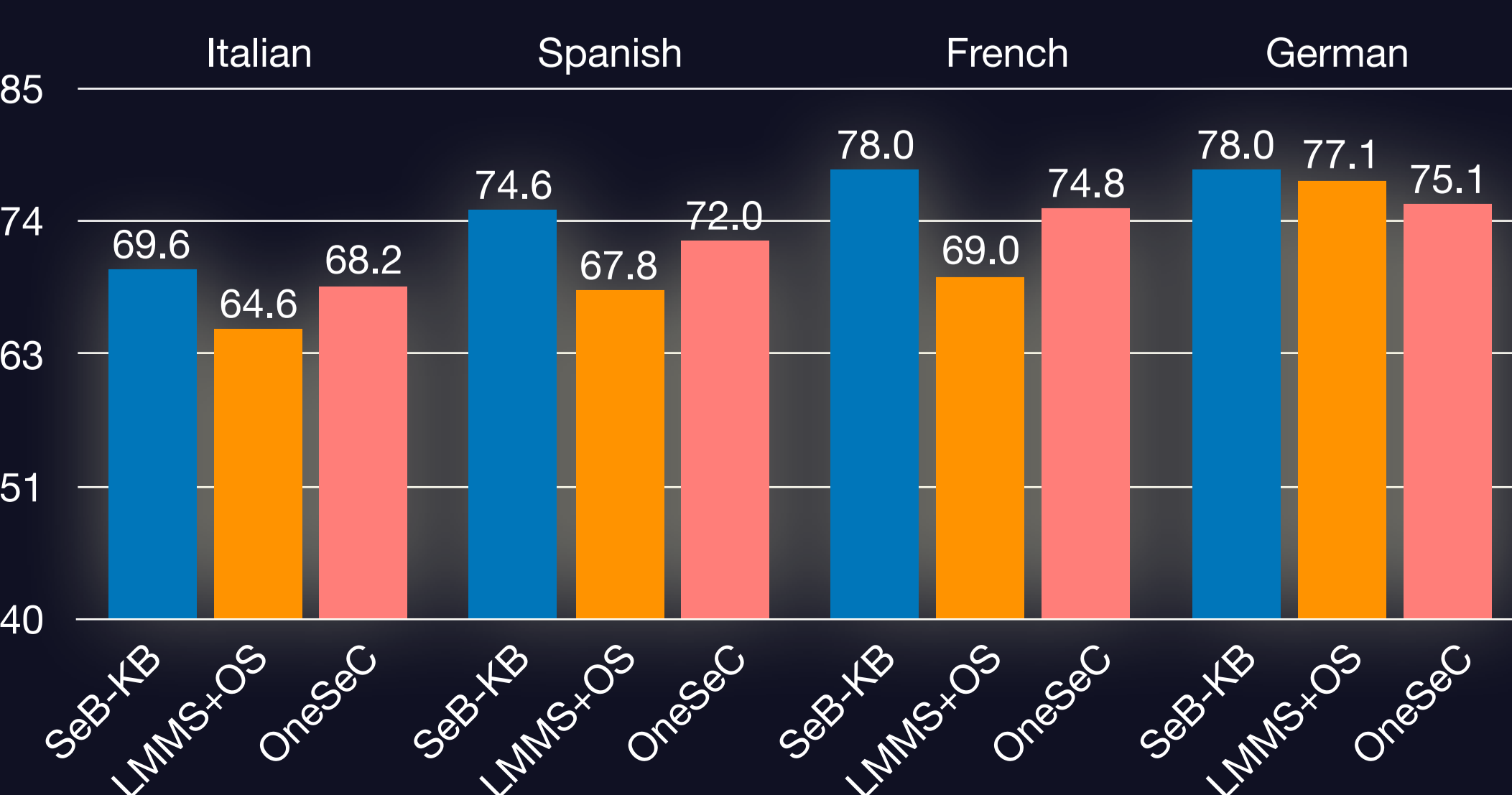Dataset containing all the instances in ALL associated with a lemma not in SemCor.

ALL - LFS

System	F1
SeB-KB	76.9
SeB-Sup	70.4
LMMS	66.7

ALL - LFW

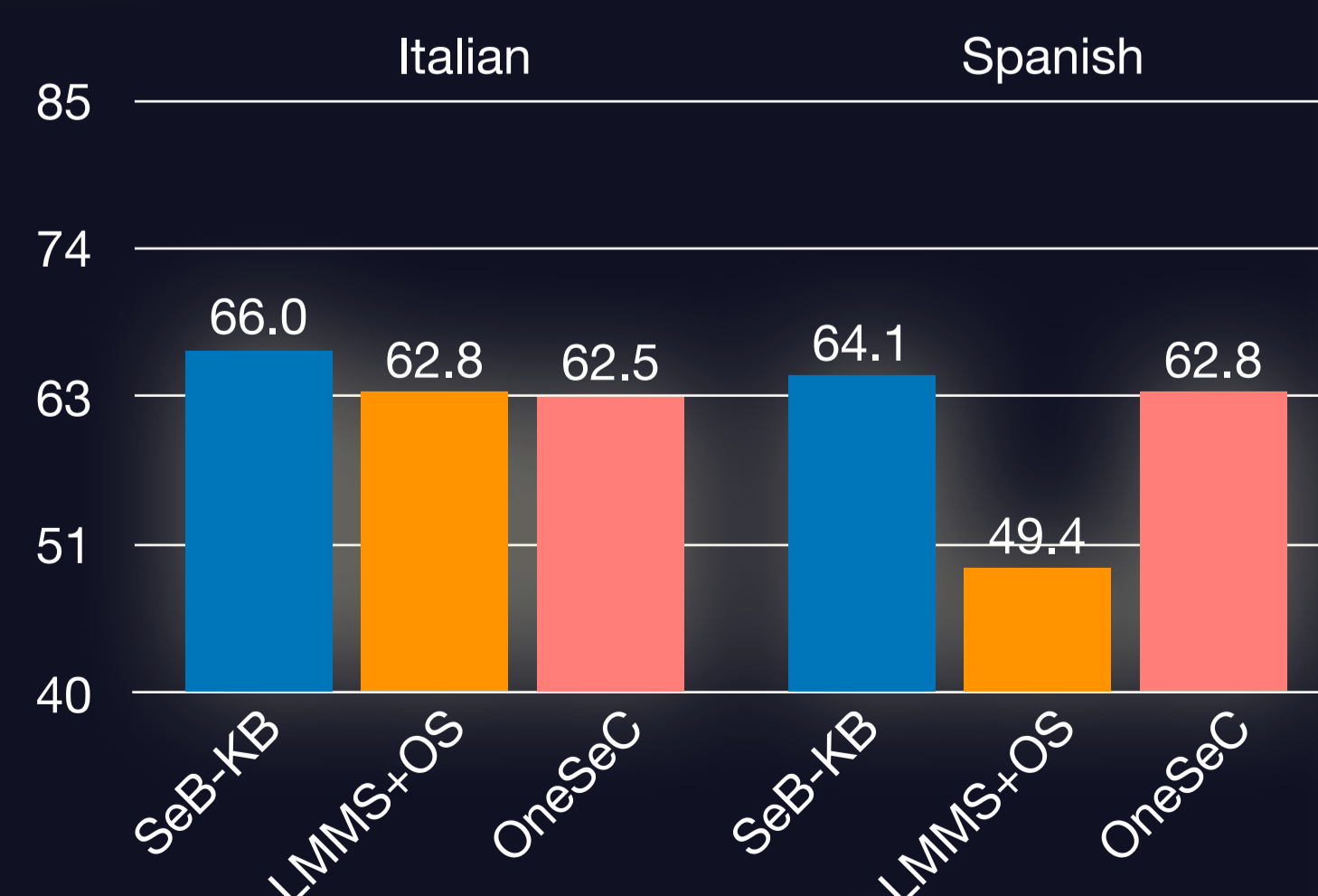| System	F1
SeB-KB	81.1
SeB-Sup	78.1
LMMS	76.3

MULTILINGUAL WSD

Competing Systems:

SeB-KB: SensEmBERT embeddings.
OneSeC: Neural supervised approach trained on OneSeC silver data.
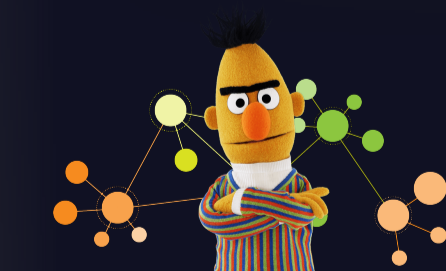**LMMS+OS**: Sense embeddings learned from OneSeC (OS).

SemEval-2013

	Italian	Spanish	French	German
SeB-KB	69.6	74.6	78.0	78.0
LMMS+OS	64.6	67.8	69.0	77.1
OneSeC	68.2	72.0	74.8	75.1

SemEval-2015

	Italian		Spanish	
SeB-KB	66.0		64.1	
LMMS+OS	62.8		49.4	
OneSeC	62.5		62.8	

SOTA everywhere!

CONCLUSIONS & FUTURE WORKS

Knowledge-based approach for producing embeddings of senses in **multiple languages**.

SOTA results on the English and multilingual WSD tasks.

NOUNS VERBS ADJECTIVES ADVERBS

Extend our approach to cover senses with **other POS tags**.