Construción dun dataset de expresións multipalabra en galego para a avaliación de modelos de lingua

Laura Castro, Anna Temerko, Marta Vázquez Abuín, Marcos Garcia

Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela
{laura.sanchez, a.temerko, martavazquez.abuin, marcos.garcia.gonzalez}@usc.gal

Introdución

- **A opacidade semántica e a polisemia** son dous dos maiores retos aos que se enfronta a tradución automática e outras tarefas relacionadas co procesamento da lingua natural (PLN).
- Estas dificultades son aínda maiores no tocante ás **expresións multipalabra** (do inglés, MWEs), e crítico no caso das linguas con menos recursos, como o galego.
- Neste traballo, presentamos un **dataset de 240 expresións multipalabra nome-adxectivo**, que presentan diferentes niveis de opacidade semántica (e.g. *exército romano, lei seca, burato negro, montaña rusa*), casos de polisemia e distintos niveis de frecuencia de uso.
- O seu obxectivo é servir de **ferramenta para avaliar como se desenvolven os modelos de lingua** no tratamento de MWEs de diferentes niveis de opacidade semántica, así como á hora de desambiguar os diferentes sentidos das expresións.

Metodoloxía

O dataset elaborouse da seguinte maneira:

- **Extraéronse** dependencias nome-adxectivo da Wikipedia en galego para obter unha escolla preliminar.
- **Ordenáronse** segundo o número de aparicións.
- **Seleccionáronse** manualmente expresións de diferentes niveis de frecuencia e, a priori, diferentes niveis de opacidade semántica, tanto dun único sentido como polisémicas.

Unha vez se obtivo unha lista de expresións multipalabra:

- **Definíronse** os posibles sentidos que cada unha delas podía adquirir en función do contexto.
- **Clasificáronse** segundo os niveis de opacidade semántica en **composicionais** (e.g. *exército romano*), **parciais** (e.g. *lei seca*), **idiomáticas** (e.g. *burato negro*), e **potencialmente idiomáticas** se se poderían clasificar con máis dun nivel dependendo do contexto (e.g. *montaña rusa*).
- **Creáronse** dúas oracións para cada sentido de cada MWE de forma manual para poñelas en contexto.
- **Realizouse** unha revisión externa das oracións para comprobar se cada par compartía o mesmo sentido.

Resultados

Elaborouse un dataset formado por:

- **240 expresións multipalabra nome-adxectivo** con diferentes niveis de opacidade semántica e diversas frecuencias de uso.

Niveis opacidade semántica	MWEs
composicionais	115
parciais	65
idiomáticas	18
poten. idiom.	42
totalis	240

- En total, defiríronse **322 sentidos**:

Niveis opacidade semántica	Sentidos MWE
composicionais	189
parciais	85
idiomáticos	48
totalis	322

- Cada MWE contextualizouse mediante **dúas oracións para cada sentido**, cun total de 644 oracións, as cales foron validadas.

Discusión

Este traballo presenta un recurso de gran valor e relevancia que facilitará a avaliación das capacidades semánticas dos modelos de lingua en galego. Co fin de completar o recurso, **establécense os seguintes obxectivos** de continuación a curto prazo:

- **Expandir os contextos** das MWEs con oracións extraídas de córpora, de forma que se posibilite avaliar se o comportamento dos modelos varía segundo o contexto sexa coñecido ou non.
- **Obter unha validación** externa das oracións extraídas dos córpora, nun proceso similar ao levado cabo para as manuais, para asegurar a calidade dos contextos.
- **Obter unha validación** externa da clasificación de niveis de opacidade semántica, co fin de afianzar a fiabilidade do recurso.

Conclusións

- Este dataset constitúe **un recurso esencial** para o estudo do tratamento das expresións multipalabra en PLN.
- No futuro, empregarase para **avaliar como procesan os modelos de lingua** as expresións multipalabra de diferentes niveis de opacidade semántica e os casos de polisemia.

Agradecementos

- Xunta de Galicia (ERDF 2014-2020: Axuda ED431G 2019/04 e ED431F 2021/01)
- MCIN/AEI/10.13039/501100011033 (proxectos PID2021-128811OA-I00 e TED2021-130295B-C33, e contrato PRE2022-102762)
- Axuda *Ramón y Cajal* (RYC2019-028473-I)