Avaluació de la Similitud Semántica

Definició de la tasca

La *Semantic Textual Similarity* es l’avaluació de la similitud del significat entre parells de frases. Aquesta similitud és important per les tecnologies de cerca, comprensió i generació automàtiques del llenguatge. L’anotació la faran dos anotadors diferents. Els parells de frases candidates han sigut extretes de la viquipedia catalana, d’un crawling dels webs del domini .cat, i de notícies de la Agencia catalana de Notícies.

Les pautes d'anotació del Semantic Textual Similarity per al català segueixen les de la tasca del SemEval 2017 i mesuren aquesta similaritat mitjançant una escala de sis graus, entre 0 i 5, i que qualsevol parlant d’una llengua pot anotar (no cal tenir coneixements lingüístics específics), segons la següent interpretació:

**5 -**The two sentences are completely equivalent, as they mean the same thing.

**Les dues frases són completament equivalents, ja que signifiquen el mateix.**

És a dir, diuen el mateix, encara que facin servir paraules diferents.

*The bird is bathing in the sink.*

*Birdie is washing itself in the water basin.*

*L’ocell es banya a l’aigüera*

*El moixó es renta a la pica*

**4 -**The two sentences are mostly equivalent, but some unimportant details differ.

**Les dues frases són força equivalents, però difereixen en alguns detalls poc importants.**

*Two boys on a couch are playing video games.*

*Two boys are playing a video game.*

*Dos nois en un sofà estan jugant a videojocs*

*Dos nois estan jugant a un videojoc*

**3 -**The two sentences are roughly equivalent, but some important information differs/missing.

**Les dues frases són aproximadament equivalents, tot i que alguna informació important és diferent o no hi és.**

*John said he is considered a witness but not a suspect.*

*“He is not a suspect anymore.” John said.*

*El John va dir que se’l considera un testimoni, i no un sospitós*

*“Ja no és un sospitós”, va dir en John*

**2 -**The two sentences are not equivalent, but share some details.

**Les dues frases no són equivalents, però comparteixen alguns detalls.**

*They flew out of the nest in groups.*

*They flew into the nest together.*

*Van volar del niu en grups*

*Van volar junts cap al niu*

**1 -**The two sentences are not equivalent, but are on the same topic.

**Les dues frases no són equivalents, però tracten del mateix tema.**

*The woman is playing the violin.*

*The young lady enjoys listening to the guitar.*

*La dona toca el violí*

*A la noia li agrada escoltar la guitarra*

**0 -**The two sentences are completely dissimilar.

**Les dues frases són completament diferents**

*The black dog is running through the snow.*

*A race car driver is driving his car through the mud.*

*El gos negre corre per la neu*

*Un pilot condueix el seu cotxe pel fang*

**Taula-Resum:**

|  |  |
| --- | --- |
| **Les dues frases són completament equivalents, ja que signifiquen el mateix.** | 5 |
| **Les dues frases són força equivalents, però difereixen en alguns detalls poc importants.** | 4 |
| **Les dues frases són aproximadament equivalents, tot i que alguna informació important és diferent o no hi és.** | 3 |
| **Les dues frases no són equivalents, però comparteixen alguns detalls.** | 2 |
| **Les dues frases no són equivalents, però tracten del mateix tema.** | 1 |
| **Les dues frases són completament diferents** | 0 |

Exemples i discusió

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Frase 1** | **Frase 2** | **sim.** | **Comentari/explicació** |
| **A** | Girona, Tarragona i Lleida, on més pugen els preus | Els preus pugen més a Girona, Tarragona i Lleida | 5 |  |
| **B** | Passat el pont anem pel primer camí a l'esquerra. | Seguim pel camí de l’esquerra un cop hem travessat el pont. | 5 | Les dues frases signifiquen el mateix |
| **C** | L’edició més internacional del Festival REC de Tarragona aixeca el teló | Tarragona presenta el Festival REC més internacional | 4 | Donen pràcticament la mateixa informació |
| **D** | El coll ha estat superat en 13 ocasions pel Tour de França | El coll del Glandon ha estat superat en 15 ocasions pel Tour de França. | 3 | Informen del mateix, però una part important de la informació és diferent |
| **E** | L índex de massa corporal és de 50 kg/m 2, i valors superiors a 30 kg/m 2 indiquen obesitat. | Valors d índex de massa corporal més grans que 30 kg/m 2 indiquen obesitat. | 3 | Informen del mateix, però en una de les frases falta una part de la informació |
| **F** | Dos lladres lliguen i colpegen un veí de Castellfollit de la Roca durant un assalt violent a casa seva | Detenen els dos homes que van lligar i colpejar un veí de Castellfollit de la Roca durant un assalt violent a casa seva | 2 | Comparteixen força detalls, però no diuen la mateixa cosa |
| **G** | Les sortides d'esquí de muntanya avançat es per a persones que ja han provat alguna vegada l'esquí de muntanya i volen aconseguir nous cims.. | Les sortides d'esquí de muntanya iniciació son per a persones que mai han provat aquest esport. | 2 | No són equivalents, però comparteixen alguns detalls (tipus de sortides d’esquí) |
| **H** | El problema era la falta d'institucions. | El problema no és la falta d'institucions, és la falta d'efectivitat de les institucions. | 2 | No són equivalents, però comparteixen alguns detalls (una matisa l’altra). |
| **I** | Per començar, talleu els tomàquets a trossos mitjans, afegiu-hi l'oli, la sal i el sucre i congeleu-ho una hora aproximadament. | Per començar, poseu en un cassó els préssecs pelats i tallats a daus i el sucre, i deixeu-ho coure durant 3 minuts. | 1 | Només tracten del mateix tema o tòpic. |
| J | El 2010, es calculava que la població de Malakal havia crescut a aproximadament 139.450 habitants | El 1359-1360 havia 381.456 habitants (20% menys de població). | 1 | No són equivalents però tracten del mateix tema general (demografia) |
| K | La pel·lícula va ser un altre cop un èxit. | La pel·lícula no va tenir èxit. | 1 | Diuen exactament el contrari, però dintre del mateix tòpic. |
| L | Tenim Papa, però no sabem qui és. | Però com que som més papistes que el Papa! | 0 | La 2a frase no parla del Papa, només el menciona en una frase feta |
| M | Segur que el senyor Tugores s'ho sap bé. | El qui la diu, segur que la diu bé. | 0 | No tenen res a veure |