**DICTIONARY-BASED FEATURES**

**isFirstWordTranslated**

*Checks whether the first word of the source term is a translation of the first word in the target term.*

**Example 1**: casual employment – priložnostna zaposlitev

* Giza++ entry: casual–priložnosten: 0.0805814
* Feature value: 1

**Example 2**: joinery – stavbno mizarstvo

* Giza++ entry: joinery–stavben: 0.0597653
* Feature value: 1

**isLastWordTranslated**

*Checks whether the last word of the source term is a translation of the last word in the target term.*

**Example 1**: casual employment – priložnostna zaposlitev

* Giza++ entry: employment–zaposlitev: 0.307655
* Feature value: 1

**Example 2**: joinery– stavbno mizarstvo

* Giza++ entry: joinery–mizarstvo : 0.243845
* Feature value: 1

**PercentageOfTranslatedWords**

*Ratio of source words that have a translation in the target term.*

**Example 1**: primary education – osnovnošolsko izobraževanje

* The Giza++ dictionary contains an entry for “education–ízobraževanje”, but not for “primary–ósnovnošolsko”.
* Feature value: 0.5

**PercentageOfNotTranslatedWords**

*Ratio of source words that do not have a translation in the target term.*

**Example 1**: primary education – osnovnošolsko izobraževanje

* The Giza++ dictionary contains an entry for “education–ízobraževanje”, but not for “primary–ósnovnošolsko”.
* Feature value: 0.5

**COGNATE-BASED FEATURES**

**longestTranslatedUnitInPercentage**

*Ratio of the longest contiguous sequence of source words which has a translation in the target term (compared to the source term length).*

**Example 1**: internal border of the EU – notranja meja EU

* Longest contiguous sequence with a Giza++ translation: “internal border–ńotranja meja”
* Formula: len(longest contiguous sequence) / len(source term)
* Feature value: 2 / 5 = 0.4

**longestNotTranslatedUnitInPercentage**

*Ratio of the longest contiguous sequence of source words which do not have a translation in the target term (compared to the source term length).*

**Example 1**: internal border of the EU – notranja meja EU

* Longest contiguous sequence without a Giza++ translation: “of the”
* Formula: len(longest contiguous sequence) / len(source term)
* Feature value: 2 / 5 = 0.4

**Longest Common Subsequence Ratio**

*Measures the longest common non-consecutive sequence of characters between two strings.*

**Example 1**: Low rent – nizka najemnina

* Longest common subsequence: e, n, whitespace character
* Formula: len(longest common subsequence) / max(len(source), len(target))
* Feature value: 3 / 15 = 0.2

**Longest Common Substring Ratio**

*Measures the longest common consecutive string (LCST) of characters that two strings have in common*

**Example 1**: practitioner of alternative medicine – izvajalec alternativne medicine

* Longest common substring: “ alternativ”
* Formula: len(longest common substring) / max(len(source), len(target))
* Feature value: 11 / 36 = 0.3056

**Needlemann-Wunsch Distance[[1]](#footnote-2)**

**Example 1**: practitioner of alternative medicine – izvajalec alternativne medicine

* Longest common substring: “ alternativ”
* Formula: len(longest common substring) / min(len(source), len(target))
* Feature value: 11 / 31 = 0.3548

**Dice similarity**

**Example 1**: practitioner of alternative medicine – izvajalec alternativne medicine

* Longest common substring: “ alternativ”
* Formula: 2 \* longest common substring / (len(source) + len(target))
* Feature value: 2 \* 11 / 67 = 0.3283

**Normalized Levensthein distance**

**Example 1**: practitioner of alternative medicine – izvajalec alternativne medicine

* Longest common substring: “ alternativ”
* Formula: 1 – Levensthein distance / max(len(source), len(target))
* Feature value: 1 – (15 / 36) = 0.5833

**COMBINED FEATURES[[2]](#footnote-3)**

**isFirstWordCovered**

*A binary feature indicating whether the first word in the source term has a translation or transliteration in the target term.*

**Example 1**: practitioner of alternative medicine – izvajalec alternativne medicine

* Giza++ entry: practitioner–izvajalec: 0.0967829
* Feature value: 1

**Example 2**: coal processing – obdelava premoga

* Giza++ entry: coal–premog: 0.854175
* Feature value: 1

**ADDITIONAL COGNATE FEATURES**

**isFirstWordCognate**

*A binary feature which returns True if the longest common consecutive string (LCST) of the first words in the source and target terms divided by the length of the longest of the two words is greater than or equal to a threshold value of 0.7 and both words are longer than 3 characters.*

**Example 1**: Klaipeda county ‒ Klaipedsko okrožje

* length of the longest common consecutive string of the two first words (*Klaiped*): 7
* length of the longest of the two first words in the terms (*Klaipedsko*): 10
* 7/10 = 0.7
* Feature value: 1

**isLastWordCognate**

*A binary feature which returns True if the longest common consecutive string (LCST) of the last words in the source and target terms divided by the length of longest of the two words is greater than or equal to a threshold value of 0.7 and both words are longer than 3 characters.*

**Example 1**: Latin America ‒ Latinska Amerika

* length of the longest common consecutive string of the two first words (*Ameri*): 5
* length of the longest of the two first words in the terms (*America*): 7
* 5/7 = 0.714
* Feature value: 1

1. See full paper for discussion on the Needleman-Wunsch Distance feature. [↑](#footnote-ref-2)
2. Only isFirstWordCovered feature is described. The others follow the principles stated above. [↑](#footnote-ref-3)