مقاربة جديدة لقياس التشابه الدّلالي للجمل العربيّة

A Novel Approach for the Measurement of the Semantic Similarity of Arabic Sentences

أ.زواوي ساميّة [1]، أ.رزاق خالدّ [2]، أ.كزار عقبة [3]

ملخّص: إنّ قياس التّشابه الدّلالي له أهميّة كبيرة في مجال معالجة اللغة الطّبيعيّة. حيث يسمح بحساب التّشابه بين المصطلحات المختلفة من أجل إجراء التّقريبات. تستخدم العديد من عمليّات البحث الأخيرة تقنيّات الويب والدّلالات اللغويّة لإيجاد درجة التّشابه بين مصطلحين من خلال مطابقة المعنى والعلاقات بينهما. لكن معظم مقاييس التّشابه الموجودة حاليّا تعاني من تدهور دقّة الحساب بسبب الاختلاف في البنيّة النّحويّة والدّلاليّة للجملتين المراد مقارنتهما. نقدّم في هذه الورقة مقاربة جديدة لقياس التّشابه الدّلالي بين الجمل العربيّة. لقد درسنا بعض الطّرق الموجودة لقياس التّشابه، ثمّ طبّقنا عمليّة توسّع جديدة، تعتمد على المعلومات الدّلاليّة المستخرجة من الأنطولوجيا العربيّة (Arabic Ontology).

لقد تمّ التّأكّد من آداء النّهج المقترح من خلال حساب ارتباط بيرسون (Pearson correlation) بين القيم المحسوبة والأحكام البشريّة. وقمنا بتقييمه على مجموعات البيانات المرجعيّة لـ SemEval-2017 و STS للجمل العربيّة. وبناءً على نتائجنا التّجريبيّة المتحصّل عليها، فإنّ عمليّة التّوسّع باستخدام الأنطولوجيا العربيّة، أعطت تحسّنًا كبيرًا من حيث دقّة التّشابه مقارنة بالطّرق الموجودة في نفس المجال.

الكلمات المفتاحيّة keywords: أنطولوجيا بالعربيّة، الجمل العربيّة، الأحكام البشريّة، التّشابه الدّلالي.

Abstract

The measure of semantic similarity is of great interest in the field of natural language processing. It allows calculating the similarity between different terms in order to perform estimations. Several recent searches use semantic web and ontology technologies to find the degree of similarity between two terms by matching the meaning and relationships between them. Most of the existing similarity measures suffer from calculation accuracy degradation due to the difference in the grammatical and semantic structure of the two sentences to be compared. In this paper, we introduce a novel approach for the measurement of semantic similarity between Arabic sentences. We examined some existing

[1] مخبر المعلوماتيّة الذّكيّة، جامعة محمد خيضر، بسكرة، s.zouaoui@univ-biskra.dz

[2] مخبر المعلوماتيّة الذّكيّة، جامعة محمد خيضر، بسكرة، k.rezeg@univ-biskra.dz

[3] مخبر المعلوماتيّة الذّكيّة، جامعة محمد خيضر، بسكرة، o.kazar@univ-biskra.dz

methods of measuring similarity. Then, we apply a new expansion process that is guided by the semantics information extracted from an Arabic ontology.

The performance of the proposed approach is confirmed through the Pearson correlation between the calculated scores and human judgments. We evaluate our approach on SemEval-2017 and STS benchmark datasets for Arabic sentences. Based on our experimental results, the expansion process, using Arabic ontology, gives a significant improvement in terms of similarity accuracy compared to existing methods in the area.

Keywords: Arabic Ontology, Arabic Sentences, Human Judgments, Semantic Similarity.

المدخل (Introduction): **يعدّ قياس التّشابه الدّلالي عمليّة مهمة للعديد من التطبيقات اللغويات التابعة للذكاء الاصطناعي. حيث يتم استخدامه بشكل خاص لمعالجة اللغة الطّبيعيّة (Natural Language Processing) وهو مجال علوم الحاسوب واللغويات التي تركز على معالجة المهام المختلفة، مثل استرجاع المعلومات، أو تحليل المشاعر أو الآراء، آلة الترجمة، تصنيف النّص أو تلخيصه، كشف الانتحال العلمي..إلخ[1]. هذه العمليّة تسمح بحساب درجة التّشابه بين المصطلحات المختلفة (الكلمات أو الجمل أو الوثائق أو المفاهيم .. الخ) لتنفيذ بعض المهام.**

في السّنوات الأخيرة، حظيت مشكلة قياس التّشابه الدّلالي المعتمد على مجموعة من الكلمات أو الجمل أو النّصوص العربيّة باهتمام متزايد، حيث تتمثل الفكرة في إيجاد طريقة يمكن من خلالها حساب القيمة الكميّة التي تمثل درجة التّشابه بين وحدات اللغة. تحقيق مثل هذا الهدف ليس بالمهمة السّهلة بالنّسبة للغة العربيّة. فالتّشابه الدّلالي هو مقياس يحسب التّشابه بناءً على تشابه معناه أو المحتوى الدّلالي على عكس التّشابه الذّي يمكن تقديره فيما يتعلق بتمثيلهم النّحوي (على سبيل المثال حسب شكل السّلسلة) [2]. اللغة العربيّة لديها ثراء كبير في الهياكل المورفولوجيّة والدّلاليّة، وتحتوي الكلمة العربيّة على أكثر من فئة معجميّة واحدة في سياقات مختلفة بحيث يمكن أن يكون لها معان مختلفة (خاصة بدون علامات التشكيل) التي تغير معنى الجملة في كل مرة[3].

من النّاحيّة الحسابيّة، يمكن تقدير التّشابه الدّلالي باستخدام الأنطولوجيات. أثبتت العديد من الأعمال المقدّمة في مجال معالجة اللغة الطّبيعيّة مثل اللغة الإنكليزيّة[4] أهميّة الأنطولوجيا لإنجاز مختلف تطبيقات المبرمجة في هذا المجال. نجد مثلا أن للغة الإنجليزيّة، العديد من الأنطولوجيات المفتوحة المصدر المثيرة للاهتمام والموجّهة لاستعمالها في مجالات المختلفة وأنواع محددة من المعلومات[5]. ولكن بالنّسبة للغة العربيّة هناك نقص في الموارد الدّلاليّة والأنطولوجيّات التي تجعل من الصّعب إجراء التجارب وبناء تطبيقات واقعيّة خاصة بهذه اللغة المستخدمة على نطاق واسع.

على الرّغم من أنّ فائدة التّشابه الدّلالي القائم على الأنطولوجيا قد أثبتت نجاحها في العديد من الدّراسات[6]، فإنّه لا يزال غير مستخدم على نطاق واسع كما هو متوقّع للّغة العربيّة. هناك أسباب مزدوجة: من ناحيّة، يعد بناء الأنطولوجيا مهمة صعبة، خاصة للغة العربيّة، بسبب ثرائها المورفولوجي والنّحوي والدّلالي؛ ومن ناحيّة أخرى، فإنّ العديد من الأنطولوجيات العربيّة الموجودة لا توفّر مصطلحات شاملة تساعد على استخراج المعلومات الدّلاليّة المطلوبة.

حاليّا، يعتبر علم الوجود (الأنطولوجيا) عنصراً مهمّاً جّدا للويب الدّلالي. في البحث، يمكننا العثور على العديد من التعريفات حول الأنطولوجيا، ولكن التعريف الأكثر شيوعًا هو ذلك الذّي قدمه[7] على أنّه "علم الوجود يحدد المصطلحات والعلاقات الأساسيّة التي تتضمن مفردات مجال يخص موضوعاً معيّناً بالإضافة إلى قواعد الجمع بين المصطلحات والعلاقات لتحديد امتدادات المفردات"

(An ontology defines the basic terms and relations comprising the vocabulary of a topic area as well as the rules for combining terms and relations to define extensions to the vocabulary.)

كما يمكن تعريفه على أنه نموذج يدعم الجانب الدّلالي من خلال تخزين مجموعة من المفاهيم والعلاقات بينها في سياق معين. إن بناء علم الوجود مهمة صعبة وتعد تحديا للباحثين والمبرجمين. وبناء الأنطولوجيا يمكن القيام به إما يدويًا أو شبه آلي أو آليا[8]. ويعتمد توطين المفاهيم والعلاقات الدّلاليّة بينهما على مسألتين: الهدف والسّياق الذّي سيتم فيه استخدام الأنطولوجيا، وخصائص اللغة.

على الرّغم من أنّ اللغة العربيّة هي واحدة من اللغات الأكثر انتشارا في العالم، إلّا أنّها تفتقر إلى التّطبيقات التي تدعم الأنطولوجيا. اللغة العربيّة هي واحدة من اللغات السّاميّة التي يتحدّث بها أكثر من 420 مليون، ويمكن قراءتها وكتابتها من اليمين إلى اليسار، وتتضمّن بعض الميزات الخاصّة مثل علامات التشكيل (Diacritics) والأوزان (schemes) التي تتطلّب معالجة خاصّة. في هذه اللغة، يوجد ثلاثة أنواع من اللغات: العربيّة الفصحى أو الكلاسيكيّة (Classical Arabic)، والعربيّة الفصحى الحديثة أو العاديّة (Modern Standard Arabic)، واللهجات العربيّة العاميّة (Colloquial Arabic Dialects)[9]. واللغة الثانيّة هي الأكثر استخدامًا في الحياة اليوميّة، خاصّة الجامعات والمدارس والإدارات والصّحف لأنّها لا تتطلّب استعمال التّشكيل. وفي هذا العمل، نحن مهتمّون بالتّعامل مع هذا النّوع لأنّ معظم الوثائق العربيّة المتاحة على الإنترنت مكتوبة بها.

في الإنترنت، هناك العديد من الخوارزميّات التي تحسب التّشابه أو المسافة بين السّلاسل الحرفيّة (String)، والتي يتمّ تنفيذها في لغة البرمجة جافّا وهي متاحة كمصدر مفتوح مثل:-**Levenshtein, Jaro : winkler, n-Gram, Q-Gram, Jaccard index, Longest Common Subsequence edit distance and cosine similarity** ، ... الخ[10]. في هذا العمل، قمنا بدراسة بعض الطّرق وهي: **Sorensen** ,Jaro-winkler **Dice, Jaccard index and cosine similarity**. لقد تمّ تطويرها لقياس مسافة والتي يجب أن تطابق أقرب ما يمكن إلى الإدراك البشري بالنّسبة للسّلسلة. ومع ذلك، فإنّ هذه الطّرق المتوفّرة لا تدعم قياس التّشابه

الدّلالي. ونحن من خلال هذا العمل، نهدف إلى توسيع هذه الطّرق باستخدام الأنطولوجيا العربيّة من أجل تحقيق قياس التّشابه الدّلالي للنّص العربي (أي تشابه بين الجمل) ثمّ نقارن بين هذه الطّرق فيما بينها.

بالنّسبة للتّمثيل الدّلالي، استخدمنا الأنطولوجيا (SchemNet) وهي مبنيّة على أساس الأوزان العربيّة ومعانيها. وهذه الأنطولوجيا، تحتوي على مزيج من محتويّات WordNet العربيّة [11] و [12] و[13] VerbNet .. ولقد لاحظنا أنّه لم يتمّ استخدام الأوزان العربيّة ومعانيها في AWN وAVN على الرّغم من أهمّيتها الكبيرة في اللغة العربيّة[14] . نقدّم في هذه الورقة مقاربة جديدة لقياس التّشابه الدّلالي بين الجمل العربيّة. لقد درسنا بعض الطّرق الموجودة لقياس التّشابه، ثمّ طبّقنا عمليّة توسّع جديدة، تعتمد على المعلومات الدّلاليّة المستخرجة من الأنطولوجيا العربيّة. علاوة على ذلك، وجدنا أنّ هذه المنهجيّة مثيرة للاهتمام لمحاكاة قدرة الإنسان على حساب التّشابه الدّلالي بين الجمل. يمكن أن يكون هذا مفيدًا في العثور على المستندات ذات الصّلة التي تستجيب لطلبات المستخدم.

يتمّ تنظيم بقيّة هذه الورقة على النّحو التّالي: يناقش القسم 2 الأعمال ذات الصّلة وفقًا لمعلومات الخلفيّة حول اللغة العربيّة. يوضّح القسم 3 النّظام المقترح، ثمّ في القسم 4 نصف 4 التّجارب المنجزة لتقييم النّهج المقترح عرض النّتائج. القسم 5 يناقش ويبرّر نتائج التّحليل. وأخيرًا، يلخّص القسم 6 الاستنتاج ويناقش بعض الاتّجاهات المستقبليّة.

2. **الأعمال السّابقة**: في هذا القسم، نحن مهتمّون فقط بالأعمال التي تهتمّ بدراسة مشكلة معالجة المستندات العربيّة. لقد أثبتت العديد من الأعمال البحثيّة نجاح توظيف الأنطولوجيا في المعالجة الدّلاليّة لتحقيق تطبيقات البرمجة اللغويّة الطّبيعيّة الأكثر دقّة مثل استرجاع المعلومات[14]، والإجابة على السّؤال[15]، وكشف الانتحال العلمي[16]، ...إلخ. تنقسم الدّراسة إلى مجالين رئيسيّين: الأنطولوجيّات والتّشابه الدّلالي.

1.2 **الأنطولوجيات**: في السّنوات الأخيرة، تمّ إجراء عدد كبير من الأبحاث حول بناء الأنطولوجيا العربيّة. تنتمي هذه الأنطولوجيّات إلى مجالات مختلفة وهي مبنيّة بأساليب متنوّعة. حيث شهدت العديد من الأعمال اهتمامًا متزايدًا ببناء علم الوجود العربي، وهي مهمّة صعبة ويمكن إجراؤها إمّا يدويًّا أو تلقائيًّا أو شبه تلقائيّ. إنّ القوّة التّعبيريّة للغة العربيّة تجعل من الصّعب استخراج المفاهيم والعلاقات بين المفاهيم، خاصّة عند استخدام الأسلوب التّلقائي[17]. وفقًا لـ [18]، يمكن تصنيف الأعمال المنجزة على الأنطولوجيا العربيّة على أساس أنّه يوجد أنطولوجيا المجال الإسلامي ، والأنطولوجيا اللغويّة، وتوليد الأنطولوجيا التّلقائي، وأنطولوجيا المتنوّعة.

يعتبر وورد-نات العربيّة (Arabic WordNet - AWN) [12,11]، المورد المعجميّ الوحيد الذّي يمثّل أنطولوجيا للغة العربيّة الفصحى الحديثة، والذي يستعمل على نطاق واسع من قبل الباحثين ومستخدمي الإنترنت في العالم العربي. وهذه الأنطولوجيا، تمّ إنشاؤها في عام 2006، ثمّ توسيعها في عام 2015 ليتمّ استعمالها لعدّة لغات. يحتوي الإصدار الحالي على 9916 مترادفات و17785 كلمة و3733 معنى. المرادفات

(Synsets) عبارة عن مجموعة من الكلمات العربيّة مع مرادفاتها وعلاقاتها الدّلاليّة. يعتمد العديد من الباحثين على AWN في تطوير مهمّتهم مثل العمل المقدّم في[19] والذّي يهدف إلى فهرسة الوثائق العربيّة وتطوير نظام استرجاع المعلومات باستخدام AWN كمورد دلالي، وتطبيق خوارزميّة Lesk يرفع اللبس أو الغموض. لقد استخدموا AWN للفهرسة الدّلاليّة للمستندات والاستعلامات معا.

ونجد أيضا العمل المقدّم في [13]، حيث قام بتصنيف مجموعة من الأفعال العربيّة وفقًا لطريقة ليفين وسمّيت بـ Arabic VebNet (AVN). تحتوي هذه الأخيرة على 336 فئة و7744 أفعال و1399 إطارًا ويتمّ فيها توفير المعلومات حول جذر الفعل، والشّكل اللفظي، والمشاركة، والأدوار المواضيعيّة، والإطارات والأوصاف النّحويّة والدّلاليّة للأفعال. كل فئة هي هيكل هرمي يوفّر معلومات نحويّة ودلاليّة حول الأفعال ويوجهّها إلى فئات فرعيّة.

عمل آخر مقدّم من طرف[20]، والذّي من خلاله يقوم بتحديد مجموعة من المفاهيم لكل كلمة عربيّة، والعلاقات الدّلاليّة بين هذه المفاهيم. لقد قام ببناء المستويات العليا من شجرة علم الوجود العربيّة، والتي تمثل المفاهيم الأكثر تجريدًا في اللغة العربيّة مع العلاقات الفلسفيّة والمنطقيّة. بينما اقترح عمل آخر يتمثل في[18] أنطولوجيا جديدة ثنائيّة اللغة بين العربيّة والإنكليزيّة. وقد قام بتجميع الكلمات العربيّة مع عدد من علاقات الارتباط الخاصّة بهم مثل المرادفات، والمتضادات، والتشعب، والاختصار، والمرادف، سميت هذه الأنطولوجيا بـ Azhary، وتحتوي على 26195 كلمة، مرتبة في 13328 مجموعة.

وجدنا أيضا عملاً آخر مقدّم في الورقة [14]، حيث قام ببناء أنطولوجيا على أساس أوزان اللغة العربيّة لمساعدة الباحثين في العديد من مهام المعالجة الطّبيعيّة اللغويّة مثل التحليل اللغوي للنصوص المكتوبة باللغة العربيّة، وحساب التّشابه الدّلالي، وما إلى ذلك. على عكس اللغات الأخرى، تتميز اللغة العربيّة باحتوائها على الصّيغ والأوزان والبناء. تتضمن الأنطولوجيا المقترحة على تصنيف الأوزان وفقًا للمعنى الدّلالي الذّي توفره للجذر المرتبط به. يتأثّر المعنى الدّلالي للأوزان بشكل كبير بعلامات التشكيل ويمكن أن يغير أحد الأوزان معناه الدّلالي بتغيير علامات التشكيل[21]. تم تصميم الأنطولوجيا حول التّسلسل الهرمي للكلمات العربيّة، ومعلوماتها المعجميّة، وشكل الأوزان والمعنى الدّلالي المرتبط بالكلمات. ولقد تمّ استخراج الأسماء من AWN، بينما تصنيف الأفعال تمّ على أساس AVN.

من جهة أخرى، طور العمل المقدّم في[22] برنامج قائم على أساس استخدام الانطولوجيا وذلك لاسترجاع المعلومات باللغة العربيّة. يتكوّن العمل من أربع وحدات رئيسيّة، وهي محلّل الاستعلام، المفهرس البحث ووحدة التّصنيف. لقد قاموا بإنشاء فهرس دلالي من خلال ربط مفاهيم الأنطولوجيا بالمستندات بما في ذلك قيمة التّعليق التّوضيحي لكلّ ارتباط، لاستخدامه في ترتيب النّتائج. إنّ الهدف من هذا العمل هو استخدام المعلومات الدّلاليّة المحفوظة في الأنطولوجيا لتصنيف الوثائق العربيّة.

نجد أيضا الورقة المقدّمة في[5]، والذي اقترح منهجًا جديدًا للتحليل الدّلالي للنّصوص العربيّة باستخدام أنطولوجيا العربيّة والمخطّطات التّصميميّة (Conceptual Graphs - CG). لقد قام المؤلّفون ببناء

أنطولوجيا عربيّة جديدة، بناءً على محتوى الموارد اللغويّة: WordNet Arabic وVerbNet. وتحتوي هذه الانطولوجيا على قائمة بالأفعال المستخرجة من AWN مع الحالات التي صيغت في شكليّة CG المقابلة لإطاراتها النّحويّة المستخرجة من AVN. لقد تمّ استخدام هذه الحالات في خطوات المطابقة النّحويّة والدّلاليّة من أجل الحفاظ على الحالة الصّحيحة.

في سياق استغلال الأنطولوجيا، يتمّ إنشاء العديد من الأنطولوجيات العربيّة لتقديم المعلومات في مختلف المجالات ذات الاهتمام من أجل تسهيل تبادل أنواع مختلفة من المعلومات بين المستخدمين. يعتمد كلّ مجال من مجالات المعرفة على المفاهيم: الكائنات والمفاهيم الوحدات الأخرى التي من المفترض أن تكون موجودة في مجال الاهتمام بالإضافة إلى العلاقات بينهم[23]. يمكن تصنيف الأنطولوجيّات العربيّة الموجودة في مجالين وفقًا للمعلومات التي يتعاملون معها: المجال الإسلامي والمجال غير الإسلامي. الأعمال التي أنجزت في إطار المجال غير الإسلامي، تتمحور حول مواضيع مختلفة في المجالات العامّة مثل السّياحة والرّياضة والقانون وما إلى ذلك. أمّا المجالات الأخرى، فإنّه يرتكز على معالجة البيانات المتعلّقة بالقرآن الكريم والحديث النّبوي الشّريف وهي مبيّنة في الدّراسة[24].

2.2 **التّشابه الدّلالي**: تُظهر البحوث حول التّشابه الدّلالي مجموعة متنوّعة من الأساليب لقياس التّشابه الدّلالي العربي بين الوحدات النّصيّة (مثل الكلمات أو الجمل أو المستندات أو المفاهيم). في الدّراسة[25]، قدّم المؤلّفون ملخّصاً حول الجهود التي بذلها الباحثون لمهمّة قياس التّشابه الدّلالي للنّص العربي. قاموا بتصنيف البحوث الموجودة على أساس التّشابه بالنّسبة للوثيقة أو للجملة أو للكلمة، ثمّ قارنوا بين هذه الأساليب. هدفنا من خلال هذه الدّراسة هو تقييم الدّرجة التي تعكس التّشابه بين معاني الوحدات التي تمّت مقارنتها فيما بينها، والتي يتمّ استخدمها من خلال الخوارزميّات ويمكنها إدارة المعلومات النّصيّة بشفافيّة من وجهة نظر رقميّة[26]. اعتمادًا على نوع الوحدات النّصيّة المستخدمة لتقدير التّشابه الدّلالي يمكن تصنيف الأعمال المتوفّرة حاليا إلى:

1.2.2 **التّشابه الدّلالي القائم على الكلمات** (Words):

هذا النّوع يقدّر التّشابه بحساب العلاقة الدّلاليّة مستوى الكلمة[27]. تحتاج الأعمال إلى التّوصل إلى طريقة حسابيّة دقيقة للعثور على التّشابه الدّلالي لهذا النّوع من الوحدة. نظرًا لأنّ العلاقة الدّلاليّة على مستوى الكلمة تتطلّب الاستكشاف، فهناك العديد من الأنواع المحتملة للعلاقات التي يمكن اعتبارها: التّسلسل الهرمي (Hierarchical)، العلائقي (Associative)، والتكافؤ (مرادف) (Equivalence)، وما إلى ذلك[28]. علاوة على ذلك، يمثل قياس التّشابه الدّلاليّة بين الكلمات أساس العديد من التّطبيقات في مختلف المجالات، وقد تمّت دراسته وتطبيقه على نطاق واسع لحساب التّشابه النّصي.

بالاستناد للمؤلّفين[1]، هناك نوعان من الطّرق لتحديد التّشابه بين الكلمات، وهي معجميّة ودلاليّة. يعني الأوّل أنّ كلمتين متشابهتين إذا كان لديهما نفس ترتيب الأحرف. والثّاني يهدف إلى قياس درجة التّشابه التي ترتبط بها الكلمتان دلاليّا. مشكلة هذا النّوع هي أنّه يتطلّب موارد لغويّة، مثل قاعدة البيانات أو

القواميس أو الأنطولوجيا، إلخ. بالإضافة إلى ذلك، فإنّ مجال حساب التّشابه الدّلالي للكلمات العربيّة هو مهمة صعبة بسبب التباين الكبير في السّمات المورفولوجيّة والنّحويّة التي تتميّز بها اللغة العربيّة، حيث يمكن أن يكون للكلمة أكثر من معنى دلالي في سياقات مختلفة عندما لا يكون لها علامات تشكيليّة محدّدة. بناءً على العمل المقدّم في[29]، تهدف الورقة إلى تقديم مجموعة بيانات باللغة العربيّة تحت اسم التّشابه الدّلالي للكلمات العربيّة (AWSS). يحتوي على قائمة أزواج الكلمات العربيّة لأكثر من 70 زوجا مع تقييم تشابه بشري لكلّ زوج، ويمكن استخدامه لتقييم أداء بعض الأعمال المستقبليّة في هذا المجال.

2.2.2 **التشابه الدلالي القائم على الجمل** (Sentences): هذا النّوع من التّشابه يعتمد على فحص العلاقات العميقة والمعجميّة والسّطحيّة بين الكلمات وخصائصها في نفس الجملة. هذا النّوع من البحث مهمّ جدًّا للعديد من تطبيقات البرمجة اللغويّة الطّبيعيّة مثل توسيع الاستعلامات لتحسين ملاءمة البحث وتعلم الآلة واسترجاع المعلومات وإدارة المعرفة وما إلى ذلك [30]. تمّ اقتراح عدد من الأساليب، خاصّة للغة الإنكليزيّة ولكن بالنّسبة للغة العربيّة فلا يوجد بسبب هيكلها اللغوي المعقّد. إنّها مهمّة معقّدة وصعبة للباحثين من أجل تطوير طريقة لقياس التّشابه الدّلالي بين الجمل العربيّة.

بالنّظر إلى البحث الخاصّ بالتّشابه الدّلالي القائم على الجملة للغة العربيّة، توجد بعض الأعمال التي ترتكز على هذه القضيّة[6]. من بين أكثر الأعمال إثارة للاهتمام، نجد العمل المقترح في الدّراسة[1]، فلقد حاولوا قياس التّشابه الدّلالي بين الجمل العربيّة القصيرة باستخدام تمثيلات تضمين الكلمات (embedding Word). تركز الفكرة على استخدام الأشعة (vectors) كتمثيلات للكلمات في الفضاء متعدد الأبعاد لالتقاط الخصائص الدّلاليّة والنّحويّة للكلمات. اقترحوا ثلاث طرق (NoWeighting، IDFWeighting Part-of-speech weighting) وتم تقييمهم وفقًا لدقتها في 750 جملة من مجموعة MSR-video [31]. واقترح أيضا [30] و[32] طرقًا لقياس التّشابه بين الجمل العربيّة. حيث قاما كلاهما ببناء مجموعة بيانات خاصة بهما للغة العربيّة الفصحى الحديثة (MSA)، والتي تمّ استخدامها في مرحلة التّقييم. إلى جانب ذلك، هذا الفرع مهمّ للغايّة في الكشف عن الانتحال العلمي. حيث يمكن للمستخدم تغيير الشّكل المعجمي والنّحوي للجملة باستخدام كلمات مترادفة، ولكن ليس المعنى، خاصة للعربيّة. علاوة على ذلك، هناك أيضًا نقص في المجموعة العربيّة ومجموعة البيانات والمعاجم التي تتعامل مع المعرفة النّحويّة الدّلاليّة، وهي ضروريّة لأي بحث متقدم في مجالات مختلفة، خاصةً في مرحلة التقييم [31].

3.2.2 **التشابه الدلالي القائم على المستندات** (Documents): يعتمد هذا النّوع على حساب المسافة بين المستندات باستخدام الكلمات والجمل. تتمثّل الفكرة في تمثيل نصيّن، T1 وT2، من خلال أشعة المفاهيم ويمكن تقييم التّشابه بينهما وفقًا لذلك باستخدام إحدى طرق قياس التّشابه مثل: Cosine أو Jaccard أو Dice [33]. في العمل [34]، يقترح المؤلّفون طريقة للكشف عن العبارات المعادلة وتحليل التّشابه الدّلالي لأخبار

التّغريدات العربيّة. يستخدم هذا النّهج ميزات محاذاة الكلمات ويستخدم مجموعة من الميزات المستخرجة بناءً على الحساب المعجمي والنّحوي والدّلالي للكشف عن مستوى التّشابه بين أزواج التّغريدات.

على مدى العقود القليلة الماضيّة، ركّزت الأبحاث على بناء مجموعة بيانات جيدة التنظيم لقياس التّشابه الدّلالي بين نصين يحتويان على كلمات متشابهة. في هذا السّياق، فإن الهدف الرّئيسي من العمل المعروض في [6] هو بناء مجموعة للغة العربيّة وعرض أثرها في تحديد إعادة الصّياغة. في الوقت الحاضر لم نعثر على أيّة منشورات في الأدبيات تتناول مسألة قياس التّشابه الدّلالي بين الوثائق العربيّة باستخدام الانطولوجيا. لا يزال العثور على أوجه التّشابه بين المستندات أو النّصوص العربيّة أحد أهم التحديات التي تواجه معالجة اللغة الطّبيعيّة. ومع ذلك، وجدنا بعض الأعمال التي تتعامل مع التّشابه الدّلالي بين وثيقتين عربيتين. في العمل [35]، اقترح المؤلّفون نظامًا يمكنه إيجاد تشابه بين نصين عربيين باستخدام مقاييس تقنيات التّشابه الهجين. قاموا ببناء SemanticNet ، وهو يعتمد على الطّريقة المنطقيّة لتخزين الكلمات الرّئيسيّة العربيّة لمجال معين (علوم الحوسبة) ، باستخدام نفس مفهوم WordNet. تُستخدم هذه الشّبكة للعثور على أوجه التّشابه الدّلاليّة بين الكلمات وفقًا لمعادلات محددة. في [36] ، تقترح الورقة طريقة قائمة على المحتوى لتحليل تشابه الوثيقة المكتوبة باللغة العربيّة على أساس نمذجة العلاقة بين المستندات وعبارات n-gram الخاصّة بها. استخدم المؤلف تقنيات البرمجة اللغويّة الطّبيعيّة المختلفة للمعالجة المسبقة وفهرسة المستندات. تمت مقارنة الطّرق المقترحة بـ Plagiarism-Checker-X، وأثبتت النّتائج تفوقها عليها.

مع ذلك، يعتبر قياس التّشابه الدّلالي للوثائق العربيّة أحد التحديات الرّئيسيّة لأن الأساليب التي تم تطويرها حتى الآن لا تزال غير كافيّة للكشف عن التّشابه مثل البشر.

4.2.2 **التشابه الدلالي القائم على المفاهيم** (Concepts): هو يحسب المسافة بين المصطلحات باستخدام الأنطولوجيا أو WordNet. في الأدبيات، تم اقتراح العديد من النّظريات لقياس التّشابه الدّلالي على أساس الأنطولوجيا أو WorldNet بين مفهومين. تنقسم هذه المقاييس إلى أربع فئات: المقاييس المستندة إلى المسار مقاييس محتوى المعلومات، المقاييس المستندة إلى الميزات والتدابير المختلطة [37]. تقدر المقاييس القائمة على علم الوجود التّشابه بين مفهومين وفقًا لتمثيل المعرفة المنظمة، والمنطق المنطقي الذّي تقدمه الأنطولوجيا يقدم [38] و [39] قياسًا للتشابه الدّلالي بناءً على بنيّة الأنطولوجيا. استخدموا قياس التّشابه بين المفاهيم والذّي يعتمد على مسافة المسار التي تفصل بين المفاهيم.

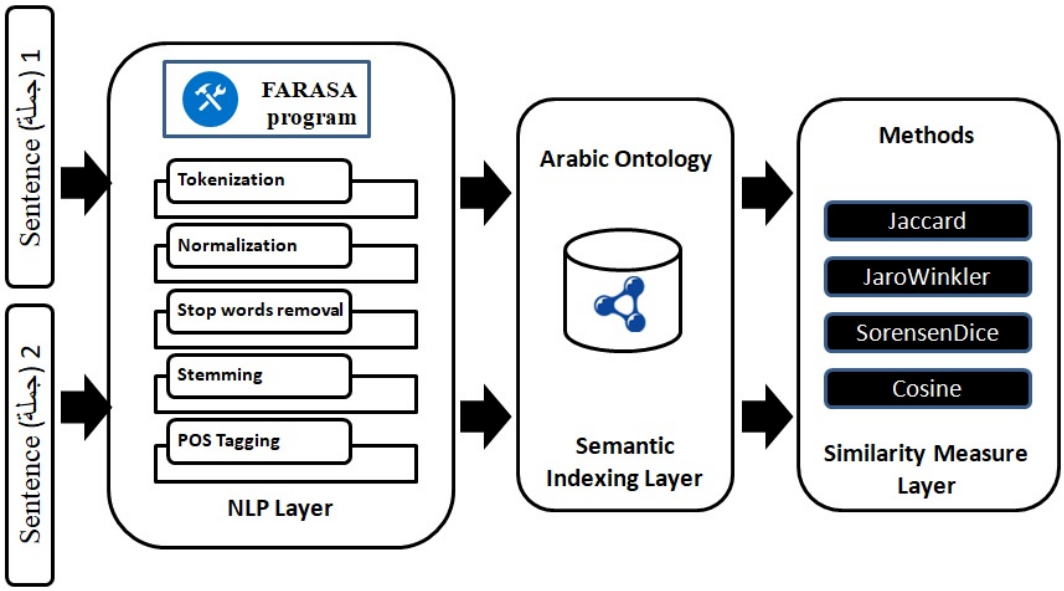
ومع ذلك، يوجد القليل من الدّراسات الحديثة التي اهتمّت بتطوير التقنيات التي تساعد في بناء الانطولوجيا بحيث تمثل المعرفة العربيّة بطريقة دلاليّة كمجموعة من مفاهيم وعلاقات[5]. فيما يتعلق بالأعمال البحثيّة في علم الوجود أو الانطولوجيا للغة العربيّة، نجد أن هناك نقصا في تطوير قياس التّشابه على أساس الانطولوجيا مقارنة بالأعمال الخاصّة باللغة الإنجليزيّة، والتي هذه الأخيرة استفادت بالفعل من البحث المكثف في هذا المجال. هناك بعض الجوانب التي تبطئ التقدم بالنّسبة للغة العربيّة في إطار البرمجة

اللغويّة الطّبيعيّة مقارنة بالإنجازات في اللغة الإنجليزيّة واللغات اللاتينيّة الأخرى مثل عدم وجود علامات التشكيل في النّص المكتوب، مما يخلق الغموض والتعقيد في القواعد النّحويّة للغة[30].

ومع ذلك، نستنتج من خلال البحوث المدروسة أن الفهرسة الدّلاليّة هي أساس العديد من تطبيقات معالجة اللغة التلقائيّة مثل استرجاع المعلومات (IR)، الإجابة على السّؤال (QA)، إلخ. بالنّسبة للغات اللاتينيّة، فلقد تم تطوير العديد من الأدوات من خلال تطبيق مجموعة من التّقنيات الذّكيّة للعثور على البيانات ذات الصّلة لتلبيّة احتياجات المستخدمين[5]. بالنّسبة للغة الإنجليزيّة كمثال، يغطي العديد من الأنطولوجيات المثيرة للاهتمام ما لا يقل عن 100 مجال مثل OpenCyc [40]. ولكن بالمقارنة مع اللغات اللاتينيّة، فإنّ اللغة العربيّة بعيدة كل البعد عن التطور التكنولوجي بسبب الاختلافات المورفولوجيّة، والثراء اللغوي والخصائص المعقدة لهذه اللغة. أحد اهتمامات جميع الأعمال الخاصّة بتطبيقات البرمجة اللغويّة الطّبيعيّة بالنّسبة للغة العربيّة هو أن الباحثين لا يجدون مجموعة بيانات منظمة بشكل جيد للسماح بالتحليل والمقارنة الموضوعيّة وتكون متاحة للجميع من أجل إعادة الاستخدام[6]. لا توجد بيانات مهيكلة بأحكام بشريّة مع استفساراتهم والتي يمكن استخدامها في مرحلة التقييم. بالإضافة إلى ذلك، تعاني معظم مقاييس التّشابه الحاليّة من تدهور في دقة الحساب بسبب الاختلاف في البنيّة النّحويّة والدّلاليّة للمصطلحات التي يتم مقارنتها.

3. **النظام المقترح**: في هذا البحث قدمنا مقاربة جديدة لقياس التّشابه الدّلالي بين الجمل. لقد درسنا بعض طرق القياس. ثم نقوم بتطبيق عمليّة توسّع جديدة (الفهرس الدّلالي)، تعتمد على المعلومات الدّلاليّة المستخرجة من علم الوجود العربي. يهدف نهجنا إلى قياس التّشابه الدّلالي بين جملتين عربيتين مكتوبتين باللغة العربيّة القياسيّة الحديثة (بدون علامات التشكيل) باستخدام بعض طرق المبرمجة بلغة جافا الحاليّة. ظهرت العديد من مقاييس تشابه السّلسلة والمسافة المتعلقة بتشابه النّص العربي في السّنوات القليلة الماضيّة. لسوء الحظ، هذه المقاييس لا تحسب التّشابه الدّلالي بين نصين (جمل أو سلسلة). إنهم لا يستخدمون معنى (Sense) الجملة التي تعكسها كلمات بناء الجملة. نحن نهدف إلى استغلال أنطولوجيا SchemNet للفهرسة الدّلاليّة للجمل قبل تطبيق قياس التّشابه.

يوضّح الشّكل 1 عمليّة التّشابه الدّلاليّة التي تحتوي على المستويات التّاليّة:



الشّكل 1: هيكل النّظام المقترح.

في هذا المستوى، نطبق مجموعة من عمليات البرمجة اللغويّة الطّبيعيّة على أزواج الجمل العربيّة كمرحلة ما قبل المعالجة على النّحو التالي:

- تقسيم (Tokenize) كل جملة لاستخراج الكلمات،
- حذف علامات التّشكيل والأحرف اللاتينيّة والمسافات البيضاء الإضافيّة والأحرف الخاصّة والأرقام،
- تغيير الكلمات عن طريق استبدال بعض الأحرف (حمزة) مثل أ و إ و آ ب ا،
- إزالة كلمات التوقف (Stop-words) مثل: إلى وهذا وغير ومن وعلى ... إلخ،
- استخلاص الجذر و أقسام الكلام ((Part-of-Speech (PoS) مثل: الاسم والفعل والصّفة لكلّ كلمة في الجملة باستخدام تقنيّة البرمجة اللغويّة الطّبيعيّة المقترحة من طرف البرنامج FARASA[41].

1.3 **مستوى الفهرسة الدّلاليّة**: نظرًا لأنّ الفهرسة الدّلاليّة تسمح بإعطاء المعنى الدّلالي للكلمات الأصليّة في شكل آخر، فإننا نستخدم استعلامات SPARQL لاستخراج هذه المعلومات مثل: المرادفات المتضادات... إلخ من الأنطولوجيا العربيّة SchemNet. والذي يقوم على استبدال الكلمة الأصليّة بأكثرها تشابهًا والتي لها نفس الدّور في الجملة (PoS). بهذه الطّريقة، سيكون لأزواج الجمل نفس التغييرات النّحويّة مع كلمات متشابهة دلاليا.

مثال: للتّوضيح أكثر نقدم الجملة التاليّة:

الجملة: "لبس الرّجل معطفا أسودا وذهب إلى السّهل ليرعى البقر."

بعد مرحلة المعالجة المسبقة التي تسمح لنا بحذف كلمات الغير مفيدة مثل: إلى تغيير النّص، لنحصل على فهرس يمثل الجملة.

الجدول 1 يعرض عمليّة الفهرسة الدّلاليّة، والتي تسمح بإضافة معلومات وفقًا لأقسام الكلام والمعنى الذّي ظهر في الأنطولوجيا SchemNet.

Word كَلِمة	PoS اقسَام الكَلَام	Root الجذر	Pattern وزن	Correct Meaning المعْنَى الصحيح	Relation علَاقة
لبس	فعل	لبس	فَعَلَ	إرتدَى	synonym
الرَجل	اسم	رجل	فَعْل	شخص	Hyponym
معطفًا	اسم	عطف	مِفْعَل	شيء	Hyponym
أسودًا	اسم	سود	أَفْعَل	لون	Hyponym
ذهب	فعل	ذهب	فَعَلَ	توجه	synonym
السهل	اسم	سهل	فَعْل	مكَان	Hyponym
يرع	فعل	رعَى	يَفْعَل	يحم	synonym
البقر	اسم	بقر	فَعَل	حيوَان	Hyponym

الجدول 1: مصطلحات الفهرس الخاصّة بالجملة المبيّنة في المثال 1

كملاحظة فإنّ الأنطولوجيا تمّ استخدامها لاستبدال الكلمات في الجملتين كلمة تلو الأخرى بكلمات مماثلة. على سبيل المثال: يمكننا استبدال الكلمات "إمرأة" و"طفل" و"رجل" بـ "شخص" باستخدام العلاقة Hyponym. كما تمّ استخدام الأنطولوجيا أيضا للعثور على مرادف لكلّ كلمة وفقًا لفئتها النّحويّة أو PoS (الاسم والفعل والصّفات).

2.3 **مستوى قياس التشابه**: في هذا المستوى، استخدمنا مكتبة Java التي تحسب قياس التّشابه أو المسافة بين سلسلتين مختلفتين للحروف[10]. ولقد وقع اختيارنا على بعض الطّرق التي تتطابق قدر الإمكان مع الإدراك البشري لقياس المسافة لسلسلتين. وفيما يلي الخصائص الرّئيسيّة لكل مقياس مستخدم. وننوه أن النّتيجة تعطي تقديرًا للتشابه الحسابي بين سلسلتين طول كل واحدة على التوالي m وn. جميع الأساليب المستخدمة تحدد التّشابه بين السّلاسل (0 تعني أن السّلسلتين مختلفة تمامًا و1 تعني أن السّلسلتين متطابقة). يعرض الشّكل 2 طرق التّشابه المختارة التي تم تنفيذها في Java.

```
Jaccard jaccard = new Jaccard(2);
MethodJaccard[ind]=jaccard.similarity(Sent1, Sent2);

JaroWinkler jaro = new JaroWinkler();
Methodjw[ind]=jaro.similarity(Sent1, Sent2);


SorensenDice sorensen = new SorensenDice(2);
MethodSorDice[ind]= sorensen.similarity(Sent1, Sent2);

Cosine cos = new Cosine(2);
```

info.debatty.java.stringsimilarity.Cosine

الشّكل 2: المقاييس المختارة لحساب التّشابه لسلسلتين في Java

- **Jaro-Winkler** (المرمز **Xjaro**): هو عبارة عن مسافة تحرير سلسلة تمّ تطويرها في منطقة ربط السّجلات (الكشف عن التكرارات). تمّ تصميم هذا المقياس لأنّه الأنسب لحساب التّشابه بين للسّلاسل القصيرة مثل أسماء الأشخاص واكتشاف الأخطاء المطبعيّة.
- **Cosine similarity** (المسمّى **Xcos**): وهو قياس التّشابه بين السّلسلتين على أساس جيب تمام الزاويّة بين الممثلين المتّجهين للسّلسلتين المراد مقارنتهما ويتمّ حسابه بهذه المعادلة:(|V2|*|V1|)/V1.V2
- **Jaccard index** (المسمّى **Xjaccard**): يتمّ تحويل سلاسل الإدخال أوّلاً إلى مجموعات من تسلسلات الأحرف **n** بحيث لا تؤخذ التّعداد في الاعتبار. كل سلسلة إدخال هي ببساطة مجموعة من **n-grams**. ثمّ يتمّ حساب الفهرس كالتّالي: (|V1 union V2|)/V1 inter V2
- **Sorensen-Dice coefficient** (المسمّى **Xsorensen**): مشابه **Jaccard index** ، ولكن هذه المرّة يتمّ حساب التّشابه كما يلي: $2*|V_1 \textit{ inter } V_2|/(|V_1+V_2|)$

ننوّه أنّ "طريقة + X" تعني أن طريقة حساب التّشابه بين جملتين تتمّ بعد عمليّة التوسّع (أي بعد استعمال الفهرس الدّلالي) وذلك باستخدام الأنطولوجيا العربيّة.

4. التجريب والتقييم

1.4 مجموعة البيانات

بالنّسبة للبيانات التّجريبيّة، فلقد واجهنا صعوبة كبيرة في العثور على مجموعة بيانات باللغة العربيّة لاستخدامها في هذه المرحلة التّقييميّة بسبب نقص الموارد في اللغة العربيّة مقابل اللغة الإنكليزيّة. ومع ذلك هناك مجموعة بيانات مرجعيّة لحساب التّشابه النّصي الدّلالي وتدعى: **SemEval-2017 Task 1** [42]. هذه البيانات ترتكز على تقييم قدرة الأنظمة على تحديد درجة التّشابه الدّلالي بين الجمل أحاديّة اللغة والمتقاطعة في اللغات العربيّة والإنكليزيّة والإسبانيّة. وتُعقّد المهمّة المشتركة لـ STS سنويًا منذ عام 2012

كجزء من مجموعة ورشة العمل SemEval /SEM. يتم تنظيمها في مجموعة من المسارات الفرعيّة الثانويّة ومسار أساسي واحد. في عملنا، استخدمنا البيانات المتوفرة في: Track 1، وهو ملف نصيّ يحتوي على 368 زوجًا من الجمل المكتوبة باللغة العربيّة (STS.input.track1.ar-ar.txt). يحتوي هذا الملف النّصي على أربعة حقول على النّحو التالي: ID (معرف فريد لكل زوج)، درجة STS (رقم بين 0 و5)، الجملة الأولى (S1) والجملة الثانيّة (S2) (انظر الشّكل 3). تمت ترجمة أزواج الجمل العربيّة يدويًا من الإنجليزيّة بواسطة خبير عربي. درجة STS هي متوسط قيمة التّشابه التي تم توفيرها يدويًا من قبل خمسة خبراء. إنه رقم متغير بين "0": يشير إلى أن معنى الجمل مستقل تمامًا و"5": يشير إلى المعنى متكافئ.

1	ID	STS score	S1	S2
349	MSRvid#SP608	0.4	رجل يحمل رجل آخر على ظهره.	رجل يلتقط صور لنملة.
350	MSRvid#SP609	2.8	قدمت السيدة بعض الشرائح على الروبيان.	طاهي يشرح الروبيان.
351	MSRvid#SP610	0.167	إمرأة تتدفق البيض في وعاء القلي.	رجل يلاعب اثنين من الكلاب.
352	MSRvid#SP611	0.5	إمرأة تقلي اللحم المفروم.	رجل يقطع جذع بفأس.
353	MSRvid#SP612	3.25	يتم إسقاط اللحم في وعاء.	إمرأة تضع اللحم في المقلاة.
354	MSRvid#SP613	0	إمرأة تقطع البصل الأخضر.	إمرأة تقع على قلعة الرمال.
355	MSRvid#SP615	0	آلة تبري قلم رصاص.	فتاة تركب دراجة هوائية.
356	MSRvid#SP617	0.4	رجل ينظف النوافذ.	رجل يقود سيارة.
357	MSRvid#SP618	0	شخص ما يخلط سائل.	رجل يعزف الغيتار.
358	MSRvid#SP619	1	شخص يثير الأرز.	إمرأة تكسر بيضة.
359	MSRvid#SP620	1.25	ثلاثة صبية في أزياء الكاراتيه يقاتلون.	ثلاثة رجال يمارسون حركات الكاراتيه في حقل.
360	MSRvid#SP621	0.25	رجل يضع السكين في الرذيلة.	رجل يرقص.
361	MSRvid#SP622	4.25	صبي يعزف الكمان على خشبة المسرح.	صبي على خشبة المسرح يعزف الكمان.
362	MSRvid#SP623	0.118	إمرأة تقطع رغيف وردي بسكين.	رجل يعزف الغيتار.
363	MSRvid#SP624	1	رجل يضيف شرائح لحم الخنزير إلى المقلاة.	إمرأة تقشر البطاطا.
364	MSRvid#SP625	0.4	الرجال يقاتلون بعضهم البعض.	رسم الرجل على ورقة بيضاء.
365	MSRvid#SP627	0	فتاة تقرأ صحيفة.	طاهي يقشر البطاطا.
366	MSRvid#SP628	2.5	لوريس بطيء يلدغ أصابع شخص.	حيوان صغير يمضغ على إصبع.
367	MSRvid#SP629	0.75	رجل يمشي على طول الطريق من خلال البرية.	رجل يقشر بصلة.
368	MSRvid#SP630	3	يعزف الرجل غيتاره.	رجل يغني في حين يعزف غيتاره.
369	MSRvid#SP631	1.6	يتسابق السباحون في البحيرة.	النساء السباحين يغوصون من منصة الإنطلاق.
370				

الشّكل 3: مثال على أزواج الجمل من الملف النّصي لـ (Track1)

2.4 **التقييم**: يتمّ تقييم الأداء عن طريق حساب ارتباط بيرسون (Pearson correlation) بين درجات التّشابه الدّلاليّة المعينة للآلة والأحكام البشريّة. ارتباط بيرسون هو رقم بين 0 و1 يشير إلى مدى قوّة الارتباط بين خطّين. لقد استخدمنا مثل هذا المقياس لتحديد مدى ارتباط مقاييس التّشابه الدّلاليّة بالمقارنة للأحكام البشريّة، حيث تشير القيمة 0 إلى عدم وجود علاقة خطيّة بين المقاييس، وتشير القيمة 1 إلى علاقة خطيّة إيجابيّة مثاليّة بين المقاييس. لقد حسبنا هذا المقياس باستخدام مكتبة جافا التي توفرها "org.apache.commons.math3.stat.correlation". يوضّح الشّكل 4 مثالاً لطريقة حساب ارتباط Pearson بين كل مقياس تشابه والأحكام البشريّة.

```
double CorrJaccard = new PearsonsCorrelation().correlation(Human, MethodJaccard);
System.out.println(CorrJaccard);

double CorrJaro = new PearsonsCorrelation().correlation(Human, Methodjw);
System.out.println(CorrJaro);

double CorrSor = new PearsonsCorrelation().correlation(Human, MethodSorDice);
System.out.println(CorrSor);

double CorrCos = new PearsonsCorrelation().correlation(Human, MethodCos);
System.out.println(CorrCos);
```

الشّكل 4: لقطة شاشة لحساب ارتباط بيرسون في جافا.

3.4. النّتائج والمناقشة (Results and discussion): لقد تم تنفيذ النّظام المقترح على Eclipse IDE باستخدام لغة برمجة Java لتصميم واجهة المستخدم التي تعرض عناصر النّظام كما هو موضّح في الشّكل 6. لقد استخدمنا لغة استعلام SPARQL لاستخراج البيانات من الأنطولوجيا (SchemNet).

Measuring Semantic Similarity of Arabic Sentences Using Ontology

حساب التشابه الدلالي بين جملتين باللغة العربية باستخدام الأنطولوجيا SchemNet

Num	Sent1	Sent2	Hum.	XCos	XJaccard	XJaro	XSorensen
229	جرو يكرر التقلب من جنب إلى ...	كلب يتقلب من جنب إلى جنب.	0.7	0.73	0.39	0.68	0.57
39	حيوان يأكل.	الحيوان يقر.	0.08	0.55	0.38	0.7	0.55
189	دب الباندا يمضغ العصا.	يلعب رضيع باندا بالعصا.	0.6	0.53	0.41	0.76	0.58
53	دجاجة تقر قرأ ميتا.	ديك يقر قرأ ميتا.	0.72	0.69	0.52	0.87	0.68
47	رجال يجرون في السباق.	المتسابقون يجرون في المسار.	0.52	0.57	0.38	0.75	0.55
171	رجال يلعبون كرة القدم.	فريقان يلعبان كرة القدم.	0.6	0.68	0.58	0.78	0.73
148	رجل يصب الأرز في القدر.	رجل يضع الأرز في قدر عميق.	0.84	0.67	0.5	0.79	0.67
65	رجل تم وضعه في سيارة الإسعا...	رجل يصب المعكرونة في طبق.	0.0	0.35	0.21	0.67	0.35
236	رجل حفر القويا في الخشب.	رجل يحفر الخشب.	0.84	0.68	0.5	0.83	0.67
32	رجل ذو نطّاره يتكلم.	رجل في المكتبة يتكلم.	0.65	0.43	0.3	0.76	0.46
49	رجل يقطّع الطماطم إلى قطع.	إمرأة تقطّع البطاطا.	0.36	0.32	0.21	0.7	0.34
64	رجل مسطق و يطلق النار من الم...	رجل مسطق على الأرض يطل...	0.76	0.78	0.49	0.82	0.66
111	رجل و إمرأة يسيران.	رجل وامرأة يسيران ممسكان ...	0.64	0.53	0.38	0.79	0.55
132	رجل وإمرأة يرقصان البوليرد.	رجل وإمرأة يرقصان تحت ال...	0.33	0.72	0.56	0.94	0.72
282	رجل يأخذ شريحة من البيتزا.	رجل يتحدث على هاتفه المحم...	0.16	0.27	0.16	0.62	0.27
245	رجل يأخذ قطعة من بيتزا يبرون...	الرجل أخذ قطعة من بيتزا البب...	0.55	0.8	0.62	0.68	0.76
99	رجل يأكل الخبز.	رجل يعزف المزمار.	0.16	0.44	0.26	0.8	0.41
108	رجل يأكل الخشب.	رجل يضيف الزيت إلى السياره.	0.1	0.41	0.19	0.65	0.32
228	رجل يأكل.	إمرأه تسلق الصخور.	0.0	0.0	0.0	0.54	0.0
15	رجل يتأرجح في حبل.	فرد يتأرجح في الأشجار.	0.25	0.55	0.37	0.71	0.54
251	رجل يمثل النفاق.	أضاف الرجل الفوابل إلى الماء ...	0.48	0.4	0.16	0.53	0.28
316	رجل يتحدث على الهاتف.	يمضغ دب الباندا عصا.	0.0	0.15	0.08	0.53	0.15
5	رجل يتحدث على خشبة المسرح ؟	رجل يتحدث في المنصة .	0.68	0.57	0.39	0.83	0.57

الشّكل 5: مثال على نتائج تشابه بين الجمل.

في ما يلي، سنعطي مثالاً على مقاييس التّشابه بين جملتين عربيتين مع وبدون الفهرسة الدّلاليّة.

مثال: لتكن زوج الجمل: S1 وS2 ذات المعرف: ID=MSRvid\#SP579

MSRvid\#SP579=
- **الجملة 1 (S1): النّساء يطعمن الحيوان.**
- **الجملة 2 (S2): إمرأة تطعم حيوانا باليد.**
- **التقييم البشري: 5/3 = 0.60**

1/ البحث عن أقسام الكلام لكل كلمة في كل جملة عربيّة.

أقسام الكلام
- **الجملة 1 (S1): اسم فعل اسم.**
- **الجملة 2 (S2): اسم فعل اسم اسم.**

2/ إيجاد جذر كل كلمة في كل جملة عربيّة.

جذر
- الجملة 1 (S1): امراة طعم حيوان.
- الجملة 2 (S2): امراة طعم حيوان يد.

3/ استخرج المعلومات الدّلاليّة من الأنطولوجيا (المرادفات الإحتواء .. الخ) باستخدام PoS والجذر والكلمة.

فهرسة دلاليّة
- الجملة 1 (S1): (شخص، كثير) (يغذي، يطعم) حيوان.
- الجملة 2 (S2): (شخص، واحد) (يغذي، يطعم) حيوان شيء.

4/ حساب التّشابه باستخدام أربع طرق محددة (Jaccard وJaro وSorensen وCosine)، نحصل على درجات التّشابه التاليّة:

بدون المعنى:

$$Sim\_Jacard(S1,S2) = 0.21$$

$$Sim\_Jaro(S1,S2) = 0.63$$

$$Sim\_Sorensen(S1,S2) = 0.35$$

$$Sim\_Cosine(S1,S2) = 0.37$$

بوجود المعنى:

$$Sim\_Jacard(S1,S2) = 0.5$$

$$Sim\_Jaro(S1,S2) = 0.82$$

$$Sim\_Sorensen(S1,S2) = 0.67$$

$$Sim\_Cosine(S1,S2) = 0.67$$

إذا قارنا النّتائج التي تم الحصول عليها، فإننا نلاحظ أن الحسابات بدون الفهرسة الدّلاليّة كانت بعيدة جدًا عن التقييم البشري (0.6). والعكس، بعد تطبيق الفهرسة الدّلاليّة، وجدنا أن النّتائج قد تحسنت بشكل كبير وأصبحت أقرب إلى التقييم البشري (0.6).

4.4. المقارنة: لتقييم أداء نظامنا، قمنا بمقارنته مع أعمال أخرى تستخدم نفس مجموعة البيانات. يقدم الجدول 2 مثالاً لنتائج التّشابه الدّلالي الذّي تقدمه أربع طرق (XJaccard وXJaro وXSorensen وXCos) والتي تم اختيارها من مكتبة Java[10] وثلاث طرق أخرى (No Weig, IDF, and Pos) من العمل الذّي قدمه[1] بالأحكام البشريّة (.Hum). نلاحظ أن الطّرق الأربع (XJaccard وXJaro وXSorensen وXCos) يتم استخدامها بعد الفهرسة الدّلاليّة لكل زوج من الجمل باستخدام الأنطولوجيا (SchemNet).

N	Sentence Pair	Hum.	XJaccard	XJaro	XSorensen	XCos	No Weig.	IDF	POS
1	ذهب يوسف إلى الكلية يوسف مضى مسرعا إلى الجامعة	0.90	0.84	0.91	0.9	0.91	0.71	0.78	0.82
2	إمرأة تتحدث على الهاتف صبيان يتحدثان على الهاتف	0.35	0.5	0.82	0.67	0.67	0.65	0.45	0.40
3	رجل يصب المعكرونة في طبق المتسابق في سيارة إسعاف	0.00	0.11	0.53	0.21	0.21	0.15	0.13	0.13
4	إمرأة تضع الماكياج إمرأة تضع المساحيق على وجهها	0.92	0.87	0.99	0.93	0.93	0.55	0.67	0.72
5	يزيل ترسبات السمكة رجل يزيل الترسبات من السمكة	1.00	0.93	0.9	0.96	0.96	0.85	0.92	0.94
6	كلب يقرأ كتابا للطفل يقرأ طفل كتابا عن الكلاب	0.20	0.83	0.78	0.91	0.89	0.82	0.87	0.88

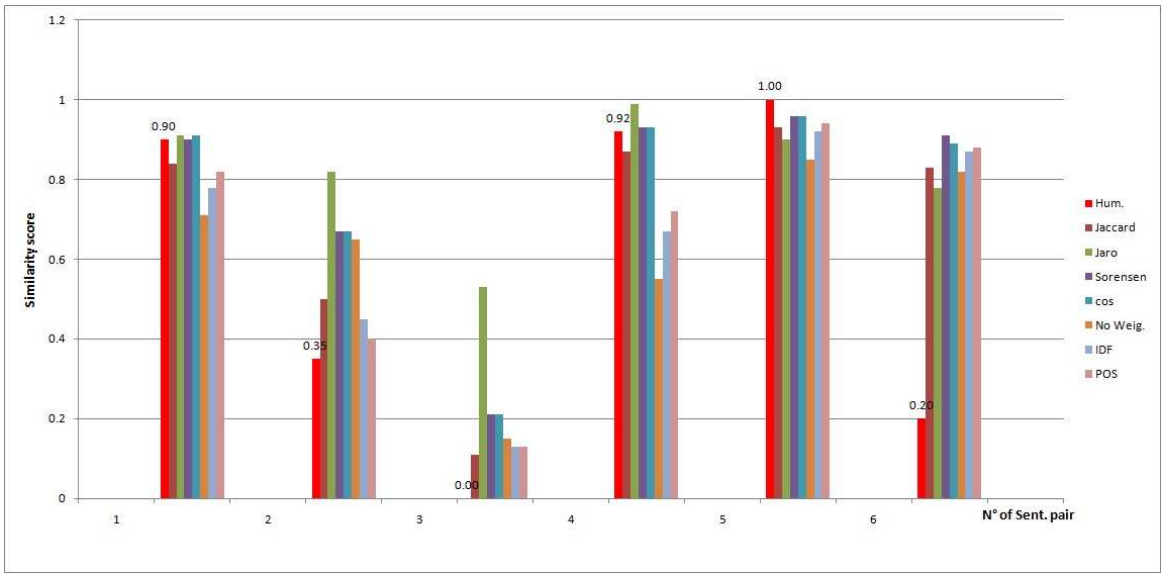
الجدول 2: مقارنة تقدير التّشابه الدّلالي المقدّمة من طرف الطّرائق المستخدمة مع الأحكام البشريّة.

تظهر النّتائج التي تم الحصول عليها في الجدول 2 أن مقاييس التّشابه XSorensenDice وXCosine تعطي بشكل عام درجات أفضل مقارنة بالأحكام البشريّة والطّرق الأخرى. ومع ذلك، في المقارنة الأخيرة (زوج الجملة رقم 6)، أعطت جميع الطّرق نتائج غير صحيحة على الرّغم من أن الجملتين تحتويان على نفس الكلمات. هذا يرجع إلى حقيقة أن ترتيب الكلمات ليس هو نفسه عند مقارنة الجملتين نحويًا.

Method	Correlation (%)
XCosine	83.31
XJaccard	82.56
XJaroWinkler	68.57
XSorensenDice	83.79
Basic method	72.33
IDF-weighting	78.20
POS tagging	79.69

الجدول 3: نتائج ارتباط بيرسون

أظهرت النّتائج المبيّنة في الجدول 3 أنّ ارتباط Pearson لـ XSorensenDice كان 83.79 ٪، وهو الأفضل مقارنة بالطّرق الأخرى. علاوة على ذلك، تؤكّد قيم الارتباط الجيّدة أهميّة استخدام الأنطولوجيا (الفهرسة الدّلاليّة) في قياس التّشابه الدّلالي بين جملتين عربيتين (كما هو موضّح في الشّكل 6).



الشّكل 6: مثال على نتائج تشابه الجمل العربيّة.

أخيرًا، نستخلص من النّهج المقترح في هذه الورقة، أن استعمال المعلومات الدّلاليّة المستخرجة من الأنطولوجيا، توفّر طريقة قياس تشابه دلالي بسيط وموثوق وفعال للغايّة.

وقد لاحظنا من خلال الطّريقة المتبعة في هذه الورقة، أنه يجب أخذ العديد من العوامل المهمة والمؤثرة عند حساب التّشابه الدّلالي وهي كالتالي:

الجملتان لهما نفس النّوع: الجملة الاسميّة أو الفعليّة. مع العلم أن الجملة العربيّة الاسميّة تبدأ بالاسم. ومع ذلك، في الجملة الفعليّة، يكون الرّأس دائمًا فعلًا.

- **الجملتان لهما نفس الكلمات بنفس الوظيفة النّحويّة وبنفس الترتيب.**
- **مع الجمل الفعليّة، يجب تحديد زمن تصريف الأفعال ما إذا كان في نفس الوقت: الماضي أو الحاضر أو المستقبل.**
- **استخدم كلمات متشابهة بمعنى كلمات أخرى وبمقصود آخر.**
- **تعديل ترتيب الكلمات التي يؤثّر على البنيّة النّحويّة للجملة.**
- **تغيير جنس الفاعل في الجملة (مذكر أو مؤنث).**
- **استخدام أدوات النّفي مثل: ليس، لم، ولا قبل الفعل وهو الأمر الذّي يمكن أن يغير معنى الجملة.**

بالمقارنة مع جميع الطّرق المذكورة في هذه الورقة، لاحظنا أن الطّريقة التي تعتمد على الأنطولوجيا تشمل حالتين:

أولاً، يمكن أن تعطي دقة أفضل لقياس التّشابه مقارنة بالطّرق الأخرى.

ثانيًا، يستغرق وقتًا طويلاً، يصل إلى أكثر من 4 دقائق لكل استعلام، وهو نقطة ضعف لهذا النّهج.

الخلاصة (conclusion): في هذه الورقة، قدّمنا نهجًا جديدًا لتوسيع بعض الطّرق الموجودة لقياس التّشابه الدّلالي بين الجمل العربيّة باستخدام الأنطولوجيا (SchemNet). حيث يعتمد هذا الأخير على الأوزان العربيّة ومعناها الدّلالي للكلمات. لقد قمنا بتطوير نظام لتطبيق عمليّة التّوسع على أربعة مقاييس تشابه (Jaccard وJaro وSorensenDice وCosine)، والتي يتمّ اختيارها من مكتبة Java. لقد جرّب النّظام على مجموعات البيانات المعياريّة SemEval-2017 للجمل العربيّة. وتمّ تقييم الأداء عن طريق حساب ارتباط بيرسون بين قيم التّشابه الدّلاليّة المعيّنة للآلة والأحكام البشريّة. استنادًا إلى درجة ارتباط بيرسون، تُظهر النّتائج أنّ مقياس التّشابه القائم على الأنطولوجيا يحقق درجة أفضل مقارنة بالطّرق الأخرى. يمكن أن يكون هذا النّهج المقترح أساسًا للعديد من تطبيقات معالجة اللغة الطّبيعيّة مثل استرجاع المعلومات (IR) والإجابة على الأسئلة (QA)، وما إلى ذلك. في العمل المستقبلي، نخطّط لدراسة تأثير القواعد التّقليديّة مثل الإعراب للنّص العربي عند قياس التّشابه الدّلالي بين جملتين.

المراجع (References)

[1] Nagoudi, E.B. and Schwab, D. (2017). Semantic Similarity of Arabic Sentences with Word Embeddings. Proceedings of The Third Arabic Natural Language Processing Workshop (WANLP), pages 1824, Valencia, Spain, April 3, 2017.

[2] Couto, F. and Lamurias A. (2018). Semantic similarity definition. Reference Module in Life Sciences (Encyclopedia of Bioinformatics and Computational Biology). doi:10.1016/B978-0-12-809633-8.20401-9.

[3] Adnen, A. and Zrigui, M. (2017). Semantic similarity analysis for paraphrase identification in arabic texts. 2017. The 31st Pacific Asia Conference on Language, Information and Computation PACLIC 31. At: University of the Philippines Cebu, Cebu, Philippines.

[4] Wali, W., Gargouri, B., and Hamadou, A. (2017). Sentence similarity computation based on wordnet and verbnet. Computacin y Sistemas, Vol. 21, No. 4, 2017, pp. 627635. doi: 10.13053/CyS-21-4- 2853.

[5] Nasri, M., Bouzoubaa, K.M. and Kabbaj, A. (2016). A novel approach for semantic analysis of Arabic texts using an Arabic ontology and Conceptual Graphs. DOI: 10.13140/RG.2.1.1493.4646.

[6] Zrigui, M. and Mahmoud, A. (2019). Similar meaning analysis for original documents identification in arabic language. In: Nguyen N., Chbeir R., Exposito E., Aniort P., Trawiski B. (eds) Computational Collective Intelligence. ICCCI 2019. Lecture Notes in Computer Science, vol 11683. Springer, Cham.

[7] Neches, R., Fikes, R. and Finin, .T, Gruber, T., Patil, R., Senator, T., and Swartout, W. (1991). Enabling technology for knowledge sharing. Arabian Journal for Science and Engineering, 1991. AI Magazine, 12(3) :36-56.

[8] Yauri, A.R., Kadir, R.A., Azreen Azman, A. and Azmi, M.A.M. (2012). Quranic-based concepts : Verse relations extraction using manchester owl syntax. International Conference on Information Retrieval & Knowledge Management.

[9] Belkredim, F. Z. and El-Sebai, A. (2009). An ontology based formalism for the arabic language using verbs and their derivatives. Communications of the IBIMA, vol. 11, pp. 44–52, 2009.

[10] Thibault Debatty (2020). Java-string-similarity. Available from https://github.com/tdebatty/java-string-similarity, last accessed on Avril 05th, 2020.

[11] Black,W., Elkateb, S., Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A. and Fellbaum, C. (2006). Introducing the arabic wordnet project. Introducing the Arabic WordNet Project, in Proceedings of the Third International WordNet Conference. Sojka, Choi, Fellbaum and Vossen eds.

[12] Elkateb, S., Black, W., Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A. and Fellbaum, C. (2006). Building a wordnet for arabic. 2006. in Proceedings of the Fifth International Conference on Language Resources and Evaluation, Genoa, Italy.

[13] Mousser, J. (2010). A large coverage verb lexicon for arabic. In: Proceedings of the 7th conference on International. Language Resources and Evaluation (LREC) (2010), Valetta, Malta.

[14] Zouaoui, S. and Rezeg, K. (2019). Ontological Approach Based on Multi-Agent System for Indexing and Filtering Arabic Documents. Journal of Digital Information Management, 17(3). 10.6025/jdim/2019/17/3/145-163.

[15] Hakkoum, A. and Raghay, S. (2016). Semantic Q&A System on the Quran. Arabian Journal for Science and Engineering, 41(12):5205–5214.

[16] Wali, W., Gargouri, B., and Hamadou, A. (2018). Using Sentence Similarity Measure for Plagiarism Detection of Arabic Documents. Springer International Publishing AG, part of Springer Nature 2018.

[17] Al-Zamil, M.G.H and Qasem, A (2014). Automatic extraction of ontological relations from arabic text. Journal of King Saud University Computer and Information Sciences, vol. 26, no. 4, pp. 462–472, 2014., 2014.

[18] Ishkewy, H., Harb, H., and Farahat, H. (2014). Azhary: An arabic lexical ontology. International journal of Web & Semantic Technology. (IJWesT), vol. 5, no. 4, pp. 71–82, 2014.

[19] Abderrahim, M.A., Dib, M., Abderrahim, M.EA. et al. (2016). Semantic indexing of Arabic texts for information retrieval system. International Journal of Speech Technology.

[20] Jarrar, M. (2011). Building a formal arabic ontology. In proceedings of the Experts Meeting on Arabic Ontologies and Semantic Networks April 26-28, 2011.

[21] Badii, E. (1993). معجم الأوزان الصّرفيّة (glossary of schemes). عالم الكتب للطباعة والنّشر والتوزيع (World of books for printing, publishing and distribution).

[22] Nazmy, T., and Elsehemy, A. (2016). Enhanced Arabic Semantic Information Retrieval system based on Arabic Text Classification. Proceedings of IMCIC - ICSIT 2016 Enhanced.

[23] Harrag, F. and Al-Nasser, A. (2014). Using association rules for ontology extraction from a quran corpus. 5th International Conference on Arabic Lan-guage Processing (CITALA 2014), Oujda, Morocco..

[24] Suryana, N. and Azmi, M. S. and Utomo, A. S. (2018). Quran Ontology : Review On Recent. Journal of Theoretical and Applied Information Technology, 96(3):568–581.

[25] Alian, M. (2018). Arabic Semantic Similarity Approaches Review. International Arab Conference on Information Technology (ACIT).

[26] Batet, M. and Snchez, D. (2014). Review on semantic similarity. In book: Encyclopedia of Information Science and Technology,

Third Edition. DOI: 10.4018/978-1-4666-5888-2.ch746.

[27] Almarsoomi, F.A., Shea, J.D.O., Bandar, Z., Crockett, K. and Member, S. (2013). Awss : An algorithm for measuring arabic word semantic similarity. IEEE International Conference on Systems, Man, and Cybernetics. DOI 10.1109/SMC.2013.92.

[28] Zhiqiang, L., Werimin, S. and Zhenhua, Y. (2009). Measuring semantic similarity between words using wikipedia. In International Conference on Web Information Systems and Mining, 2009, pages 251255. IEEE.

[29] Faaza, A., James, D., Zuhair, A. and Keeley, A. (2012). Arabic word semantic similarity. 2012. Proceedings of World Academy of Science, Engineering and Technology. No. 70.

[30] Wali, W., Gargouri, B. and Ben-hamadou, A (2015): A supervised learning to measure the semantic similarity between arabic sentences. In Computational collective intelligence (pp. 158167). Springer.

[31] MSR-video (2018). Microsoft research video corpus. https://www.microsoft.com/en-us/download/ details.aspx?id=52422, (last accessed Auguest 13,2018).

[32] Alzahrani, S. (2016). Cross-language semantic similarity of arabic-english short phrases and sentences. Journal of Computer Sciences. https://doi.org/10.3844/jcssp.2016.1.18.

[33] Alzahrani, S., Salim, N. and Abraham, A. (2012). Understanding plagiarism linguistic patterns, textual features and detection methods. IEEE Transactions on Systems, Man, and CyberneticsPart C: Applications and Reviews, vol. 42, no. 2, pp. 133-149, 2012.

[34] AL-Smadi, M., Jaradat, Z., AL-Ayyoub, M., and Jararweh, Y. (2017). Paraphrase identification and semantic text similarity analysis in Arabic news tweets using lexical, syntactic, and semantic features. Information Processing and Management, 53:640–652.

[35] Malallah, S., Qassim, A. and Alameer, A. (2017). Finding the similarity between two arabic texts. Iraqi Journal of Science, 2017, Vol. 58, No.1A, pp: 152-162 DOI:10.24996.ijs.2017.58.1A.16.

[36] Hussein, A. S. (2015). Arabic document similarity analysis using ngrams and singular value decomposition. Proceedings – International Conference on Research Challenges in Information Science, 2015- June:445–455.

[37] Aldiery, M.G. (2017). The semantic similarity measures using arabic ontology. 2017. Middle East University Amman-Jordan January, 2017.

[38] Ming Liu, Bo Lang, and Zepeng Gu (2017). Calculating semantic similarity between academic articles using topic event and ontology. CoRR, abs/1711.11508.

[39] Elavarasi, S.A. Akilandeswari, J. and Menaga, K. (2014). A survey on semantic similarity measure. International Journal of Research in Advent Technology, 2, 389-398.

[40] Witbrock, M.M C., Cabral, J. and Deoliveira, J. (2006). An introduction to the syntax and content of cyc. In Proceedings of the AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and its Applications to Knowledge Representation and Question Answering, pages 44-49, Stanford University, Stanford, California.

[41] Abdelali, A., Darwish, K., Durrani, N. and Mubarak, H. (2016). Farasa: A fast and furious segmenter for arabic. http://qatsdemo.cloudapp.net/farasa/, last accessed at 27/09/2018.

[42] SemEval (2020). Semeval-2017 tasks. Semantic Textual Similarity. Available from http://alt.qcri.org/semeval2017/task1/index.php?id=dataand-tools, last accessed on April 10th, 2020.