مقدمة

اللغة

# تعريف اللغة و مميزاتها

هو علم يبحث في اللغة من جميع جوانبها الصوتية والصرفية والنحوية اللفظية والدلالية والنفسية والاجتماعية والمعجمية والتطبيقية. وقد أطلق عليه اللغويون أسماء عديدة ؛ مثل فقه اللغة، وعلم اللسان، واللسانيات.

ويُقسم علم اللغة إلى قسمين أساسيين هما: علم اللغة النظري، وعلم اللغة التطبيقي، ويشمل علم *اللغة النظري* علم الأصوات والصوتيات وعلم اللغة التاريخي، وعلم الدلالة وعلم النحو والصرف. أما علم *اللغة التطبيقي*، فهو الآخر يشمل عدة فروع أهمها علم اللغة المقارن، وعلم اللغة التقابلي، وعلم النفس اللغوي، وعلم اللغة الاجتماعي، وتعليم اللغات، وتحليل الأخطاء.

تحدث ابن خلدون في مقدمته فعرف اللغة بأنها "علـم أن اللغـة في المتعـارف عليـه، هـي عبـارة المتكلـم عـن مقصـوده، وتلـك العبـارة فعـل لسـاني ناشئ عـن القصـد بإفـادة الـكلام، فلابـد أن تصـير ملكـة متقـررة في العضـو الفاعـل لهـا، وهـو اللسـان، وهـو في كل أمـة بحسـب اصطلاحاتها"

و جاء تعريف اللغة بحسب قاموس Merriam-Webster

"***language***:the words, their pronunciation, and the methods of combining them used and understood by a community"

يهتم علم اللغة عموماً بدراسة اللغة بوصفها نظاماً للاتصال بين البشر. ورغم أن الظاهرة اللغوية قد شغلت رجال الفكر والفلسفة منذ قرون، إلا أن علم اللغة لم يبرز علماً قائما بذاته، إلا في مرحلة متأخرة من تاريخ المعارف الإنسانية.

وظيفة اللغة الأساسية هي التواصل الاجتماعي وهو عبارة عن نشاط تفاهمي يحدث بين أفراد المجتمع الإنساني حيث ما حلوا. و لذلك فإنه عندما يبتدر شخص ما الحديث فهو يهدف الى خلق تواصل مع الآخرين. و هذا يعنى أن استخدام اللغة بين أفراد المجتمع الإنسانى يقوم على وظيفة أساسية هي : التواصل الكلامي الذي يؤديه كل من المتحدث والسامع.

اكتمال التواصل بين كل من طرفي الكلام يتوقف على كل طرف في نجاح الدور الذي يقوم به. و ذلك لأن تحقيق الدقة في الكلام و أدائه بالوجه الذي يكتمل به الاتصال والتفاهم اللغوي بين المتكلمين والذي هو الغرض المنشود.

ومما سبق نجد أن اللغة هي عبارة عن نسقٍ من الرموز والإشارات التي تشكل في النهاية **إحدى أدوات المعرفة**، هذا وتعتبر اللغة إحدى أهمّ وسائل الاحتكاك والتفاهم والتواصل في شتّى ميادين الحياة بين الأفراد في المجتمع، وبدونها **يتعذّر النشاط المعرفي** للأفراد.   
**ترتبط أيّ لغة في العالم بالتفكير ارتباطاً وثيقاً**، حيث إنّ الأفكار البشريّة يمكن صياغتها دوماً عن طريق قالبٍ لغويّ حتى في حالة التفكير الداخلي أو الباطني.

إنّ تعريف اللغة بحقيقتها وعلاقتها بالإنسان تختلف عن تعريفها بوظيفتها، حيث إنّ اللغة هي الإنسان والوطن الأوّل، واللغة هي ناتج التفكير الإنساني، وهي ما يميّز القدرة الإنسانيّة عن الحيوانية، حيث إنّها ثمرة العقل. علماء النفس عرّفوا اللغة على أنّها مجموعة من الإشارات الصالحة للتعبير عن حالات الإنسان الفكريّة، والإراديّة، والعاطفيّة (الشعور)، أو أنّها الوسيلة التي يمكن من خلالها تحليل الصور، والأفكار الذهنيّة إلى خصائصها أو أجزائها، والتي تمكّن من تركيب الصورة أو الفكرة مجدداً في أذهاننا وأذهان من حولنا، وذلك من خلال تأليف كلماتٍ وترتيبها في وضعٍ خاص.

وكذلك فإن للّغة علاقة وطيدة بالثّقافة والهويّة الخاصّة بالشّعوب، فهي التي تُعبّر عن تفكير الأمم، والوسيلة الأولى في نشر ثقافات الأمم المُختلفة حول العالم، وبالتالي فإن لغة أي أمة هي التي تُشكّل هويّة الأمّة الثقافيّة التي تُميّزها عن باقي الأمم.

## أقسام اللغات

تقسم اللغات حول العالم إلى عائلاتٍ لغويّة كاللّغات الهنديّة الأوروبيّة واللّغات الإفريقيّة الآسيويّة، حيث لكلٌ منها عددٍ من اللغات ذات الخصائص والأصول المتشابهة.   
يوجد في عالمنا ما يزيد عن ستة آلاف لغة مستخدمة حول العالم، وهناك ما نسبته الثلاثين بالمئة من تلك اللغات لا يتجاوز عدد الناطقين بها الألف شخص، **ومن أهمّ اللغات على وجه الأرض** وفقاً لعددٍ من **المعايير كأهمّيتها وتأثيرها في التعاملات التجاريّة، ومكانتها العالميّة كواحدة من لغات التواصل المشترك**: اللغة الإنجليزيّة، والعربيّة، والإيطاليّة، والفرنسيّة، واليابانيّة، والصينيّة، والكوريّة، والكانتونيّة، والتايلنديّة، والبنجابيّة، والهنديّة، والجاويّة، البولنديّة، والتركيّة، والفيتناميّة، والفارسيّة، وغيرها. تتنوّع اللغات وتتطوّر مع مضي الوقت والزمان، ويمكن العودة إلى تاريخ تطوّرها وبنائها عن طريق المقارنة بين اللغات الحديثة، بهدف تحديد سمات اللغة القديمة (لغة الأجداد)، وذلك من أجل فهم المراحل التنمويّة التي يمكن حدوثها في المستقبل.

## خصائص اللغة:

1. **اللغة نظام**: فهي تقوم على قواعد، وأسس.
2. اللغة صوتية.
3. **اللغة عرفية:** للغة قواعد متعارف عليها والعرف هو الذي يحكمها، ولايحكمها المنطق، وعندما يتعارض المنطق والعرف يؤخذ بالعرف ويترك المنطق .
4. **اللغة إنسانية:** إن الإنسان خص باللغة من دون الأحياء.
5. **اللغة اجتماعية:** فلا يمكن للإنسان أن يتكلم من دون أن يعيش مع آخرين.
6. **اللغه سلوك مكتسب:** اللغة لا تورث إنما تكتسب.
7. **اللغة متطورة:** اللغة تنمو وتتطور، وتزداد مفرداتها، وتقبل مفردات جديدة، وقد تندثر منها مفردات فلا تستعمل في الكلام، وهي تعكس تطور الفرد، وتطور الأمة.
8. **اللغة رمزية:** الأصوات رموز. أي بدائل تستخدمها في الإشارة إلى شيء ليس هو الأصوات نفسها. فالأصوات التي تمثل الكلمة "وردة" هي وسائل للإشارة إلى الوردة نفسها وليس ثمة علاقة ضرورية أو طبيعية بين أصوات هذه الكلمة والوردة نفسها.
9. اللغة معنى.

# اللغة العربية

تنتمي اللغة العربية إلى أسرة اللغات السامية، المتفرعة من مجموعة اللغات الأفرو-آسيوية. وتضم المجموعة السامية الرئيسية، لغات حضارة الهلال الخصيب القديمة مثل الأكادية الكنعانية، والآرامية، واللغة العبرية، واللغات العربية الجنوبية، وبعض لغات القرن الأفريقي. وعلى وجه التحديد **فإن اللغة العربية تصنف ضمن المجموعة السامية** الوسطى، فتكون بذلك من ضمن اللغات السامية الشمالية الغربية والتي تشمل الآرامية والعبرية والكنعانية ، وهي أقرب اللغات السامية للعربية (جواد على، ۱۹۸۰م).

نشأت اللغة العربية الفصيحة في شمالي الجزيرة العربية. ويرجع أصلها إلى العربية الشمالية القديمة التي كان يتكلم بها العدنانيون. وهي لغة تختلف في كثير من مكوناتها وأساليبها وأصواتها عن العربية الجنوبية القديمة، التي نشأت في جنوبي الجزيرة وعرفت قديماً باللغة الحميرية، وكان يتكلم بها القحطانيون.

وتدل الشواهد على أن اللغة العربية، هي الأقرب إلى اللغة السامية الأم، التي انبثقت منها اللغات السامية الأخرى. ويرى حنا الفاخوري (1974م) صاحب تاريخ الأدب العربي أن العربية، ولاحتباسها في جزيرة العرب، لم تتعرض لما تعرضت له باقي اللغات السامية الأخرى من اختلاط، فظلت بذلك محافظة على نقائها وأصالتها، وحافظت على كل خصائص اللغة السامية الأم.

إضافة لما ذكر، فهناك العديد من الآراء والروايات حول أصل اللغة العربية إذ يذهب البعض (وهو قول راجح ومعتبر) إلى أن يعرب بن كنعان هو أول من أعرب في كلامه، وتكلم بهذا اللسان العربي فسميت العربية باسمه.

**تصنّف اللغة العربية في شكليها الكلاسيكيّ والحديث ضمن اللغات السامية الرئيسيّة**، وقد ظهرت لأوّل مرّة في شمال الجزيرة العربية وبلاد ما بين النهرين خلال العصر الحديديّ، وفي أوائل القرن الحادي والعشرين كانت اللغة العربية هي اللغة الأكثر أهمية من حيث عدد المتحدّثين بها، واللغة العربية الفصحى هي اللغة الأولى التي تحدّث بها أكثر من 200 مليون شخص يعيشون في ساحل المحيط الأطلسيّ في شمال أفريقيا وصولاً إلى غرب إيران. كما يوجد جزء كبير من الكلمات العربية التي تقع ضمن قاموس اللغات الغربية نتيجة التأثير العربيّ على أوروبا خلال القرون الوسطى.

واللغـة العربيـة أكـبر لغـات السـامية مـن حيـث عـدد المتحدثـين، وإحدى أكثر اللغات انتشـارا في العالم، يتحدثها أكثر من أربعمائة واثنين وعشـرين مليـون نسـمة، ويتـوزع متحدثوهـا في الوطـن العـربي، بالإضافـة إلى العديـد مـن المناطـق الأخـرى المجاورة والـدول الإسلامية، وبسبب أهمية اللغة العربية تم اعتبارها إحدى اللغات الرسمية لمنظمة الأمم المتحدة و جميع المنظمات الأخرى التابعة لها.

نالت اللغة العربية منذ القديم و لغاية وقتنا الحاضر اهتماما كبيرا من الباحثين.

## خصائص اللغة العربية

اللغـة العربيـة لغـة إنسـانية حيـة، لهـا نظامهـا الصـوتي والصـرفي والنحـوي والتركيـبي، كمـا لألفاظهـا دلالاتها الخاصـة بها. وقـد رأى العلمـاء أن كل خـروج عـن هـذا النظـام اللغـوي المتكامـل يعـد **لحنـا**، سـواء أكان هـذا الخـروج بخلـط الـكلام بلغـة أخـرى، أم باسـتعمال اللفظـة في غـير موضعهـا، أم في مخالفـة أي عنصـر أساسـي مـن عناصـر كيانهـا اللغـوي الـذي يميزهـا عـن غيرهـا مـن اللغـات الإنسـانية.

## مزايا وخصائص العربية

تُعدّ اللّغة العربيّة لغةً خالدةً، ولن تنقرض مع مرور الزّمن أبداً حسب دراسة لجامعة برمنجهام أُجريت للبحث في بقاء اللّغات من عدمه في المُستقبل، وتتميّز اللّغة العربيّة بالكثير من الميزات التي تميزها عن الكثير من اللّغات، **ومن هذه الميزات**:

1. **الفصاحة:** وهي أن يخلو الكلام ممّا يشوبه من تنافرٍ بالكلمات، وضعف التّأليف، والتّعقيد اللفظيّ.
2. **التّرادف:** وهو أن يدلّ عددٌ من الكلمات على نفس المَعنى المراد.
3. **الأصوات ودلالتها على المعاني:** بمعنى أن يفهم معنى الكلمة بشكلٍ عامّ أو دقيق من خلال الصّوت فقط، وهذه من أهمّ الميزات الخاصّة باللّغة العربيّة.
4. **كثرة المُفردات:** تزخر اللّغة العربيّة بعددٍ وافرٍ جدّاً من المُفردات، ولا تحتوي لغةٌ أخرى على عدد أكثر أو يُساوي العدد الذي تحتويه اللغة العربية.
5. **الثّبات:** من أكبر التّحديات التي واجهتها العربيّة هو ثباتها وانتصارها على عامل الزّمن والتطوّر، في حين أنّ اللّغات الأخرى مثل الإنجليزية قد تطوّرت واختلفت بشكل كبير عبر الزّمن.
6. **التّخفيف:** وهو أن أغلب المُفردات في اللّغة العربيّة أصلها ثلاثيّ، ثم يأتي الأصل الرباعيّ، ثمّ الخُماسيّ على التّرتيب في كثرة انتشاره في أصول المُفردات العربيّة.
7. **القـدرة علـى التوليـد:** فهـي لغـة اشتقـاق ومـع أن هـذه الظاهـرة موجــودة فـي بعض اللغـات الأخـرى إلاّ أنها في العربية أوسع وأغنـى ويعد الاشتقاق أكبر مصدر لثـراء اللغــة العربيــة وتطويعهـا لاسـتيعاب كثيـر مـن المستحدثات والمعاني الجديدة.
8. **معربة:** للّغة العربية قواعدها في تنظـيم الجملـة وفـي ضبط أواخر الكلمات بهـا ضبطاً خاصاً، وتستخدم الحركـات فـي اللغة العربية للتعبير عـن المعنى، فبواسطة الحركـات الإعرابيـة يمكن التمييـز بيـن اللفظ ومدلوله، كما يمكن رفع اللبس الذي يحيط أحيانـاً بتلـك الألفـاظ ودلالاتها، وبواسطـة الحركـة الإعرابية يمكن إبراز الصلة النحوية بين الكلمـة والأخـرى في الجملة الواحدة.

لقد تميزت اللغة العربية عن غيرها من لغات البشر بكثير من الخصائص التي جعلتها وعاء لحضارة الجنس البشري ومسايرةً ومتسعةً للثقافة العالمية المعاصـرة والتطـور التكنولوجي والتي منحتها السبق والأفضلية دون غيرها ويمكن أن يستفاد من تلك الخصائص عند تأليف وتصميم برامج خاصة باللغة العربية بحيث توظف تلك الخصائص والميـزات فـي تلك المناهج بما يتناسب مع طبيعتها وأهميتها.

# اللغة الإنكليزية

اللّغة الإنجليزيّة هي مِنَ اللّغات الدوليّة التي تنتشرُ في كافّة دول العالم تقريباً؛ بسبب استخدامها في الأعمال والدّراسة، وهي واحدة من وسائل التّواصل العالميّ والدبلوماسيّ بين الشّعوب في العالم. تُعرفُ اللّغة الإنجليزيةّ بأنّها لغةٌ غربيّة تتبعُ للّغات الجرمانيّة القديمة، والتي تطوّرت مع مرور الوقت بسبب استخدامِ العديد من اللّهجات المحكيّة معها؛ ممّا ساهمَ في إضافةِ مجموعةٍ من المفردات والتّراكيب اللغويّة الجديدة لها، وما زالت اللّغة الإنجليزيّة تتطوّرُ حتّى هذا الوقت بالتّزامنِ مع التطوّرات العالميّة في العديدِ من المجالات المُتنوّعة.

تصنّفُ اللّغة الإنجليزيّة كإحدى اللّغات الأكثر انتشاراً، وتأتي في المرتبة الثانيّة مُباشرةً بعد لغة الماندرين من حيثُ عدد المُتكلّمين بها؛ سواءً كلغةٍ أساسيّة (اللّغة الأم)، أو لغةٍ ثانويّة تُستخدمُ في أغلب المجالات العامّة.

تُعتبرُ اللّغة الإنجليزيّة اللّغة الرسميّة للعديدِ من أنواع المجالات الدراسيّة الجامعيّة؛ كالعلومِ، والطبّ، والهندسة، والاقتصاد، وإدارة الأعمال، وغيرها من المجالات التعليميّة الأُخرى التي يسعى العديدُ من الطّلاب الجامعيين لدراستها. تُساعدُ اللّغةُ الإنجليزيّة في زيادةِ القدرة على قراءةِ الكُتُب، والمجلات، والأبحاث المنشورة بالاعتماد عليها، والتي تدعمُ تنمية العلم والمعرفة والثّقافة الشخصيّة عند الأفراد.

تقدّمُ اللّغةُ الإنجليزيّة مجموعةً من وسائل الاتّصال بين الشّعوب المُختلفة، ممّا يُساهمُ في التعرّفِ على العادات، والتّراث، والتّقاليد الشعبيّة العالميّة. تُستخدمُ اللّغةُ الإنجليزيّة بشكلٍ كبيرٍ جدّاً ضمن المواقع الإلكترونيّة على شبكة الإنترنت؛ حيث إنّ تعلّمها يُساهمُ في مُساعدةِ الأفراد في التعرّف على طبيعةِ المواقع الإلكترونيّة، والمُحتويات المنشورة في صفحاتها المُختلفة.

## تاريخ اللّغة الإنجليزيّة

تعودُ بداية ظهور اللغة الإنجليزيّة إلى وصول القبائل الجرمانيّة لأراضي المملكة المتحدة في القرن الخامس للميلاد، ولقد اختلطتْ لغتهم مع اللغات السائدة في تلك الأراضيّ والمناطق؛ مما أدى إلى ظهورِ لغةٍ جديدة أُطلقَ عليها لاحقاً مُسمى اللغة الإنجليزيّة القديمة، وفي القرن الرابع عشر للميلاد تطورتْ اللغة الإنجليزيّة؛ بسبب اختلاطها مع اللغة الفرنسيّة التي أثرتْ عليها بعد احتلال القوات الفرنسيّة للمناطقِ الإنجليزيّة.

في القرن السّادس عشر للميلاد بدأ الانتشارُ الأول للّغة الإنجليزيّة خارج حدود أراضي إنجلترا، وخصوصاً مع بدايةِ تطوّر العلاقات السياسيّة والاقتصاديّة بين الإنجليز والدُّول والشّعوب المحيطة بهم، وساهم ظهور عصر الطّباعة في ذلك الوقت إلى انتشارِ العديد من المعاجم اللغويّة، والقواميس المُؤلَّفة وفقاً للّغةِ الإنجليزيّة، والتي ساعدتْ في تعليم وتعريف العديد من الأشخاص غير الإنجليز على هذه اللّغة، أمّا اللّغة الإنجليزيّة الحديثة فقد ظهرتْ منذُ القرن التّاسع عشر للميلاد، والتي أصبحتْ لغةً رسميّةً وثانويّةً عند العديد من دول العالم، وما زالت اللّغة الإنجليزيّة تشهدُ تطوّراً مُستمرّاً حتّى هذا اليوم.

## خصائص اللغة الإنجليزيّة

تتميّزُ اللّغة الإنجليزيّة بمجموعة من الخصائص التي تُميّزها عن اللّغات العالميّة الأُخرى؛ وهي:

1. تحتوي اللّغةُ الإنجليزيّة على عددٍ كبيرٍ من التّراكيب والمفردات، والتي يصلُ عددها الإجماليّ إلى أكثر من 600,000 كلمة.

2. اختلطتْ اللّغةُ الإنجليزيّة مع العديد من اللّغات العالميّة، ومن أهمّها اللّغة العربيّة، واللّغة الفرنسيّة، واللّغة اللاتينيّة، واللّغة الألمانيّة، واللّغة الإغريقيّة، وغيرها من اللّغات الأُخرى.

3. يتشابه نطقُ معظم الكلمات الإنجليزيّة معاً، ويتطلّبُ ذلك من الشّخص الذي يتعلّمُها مزيداً من الوعي، والتّركيز في فهم هذه الكلمات؛ حتى يتمكّن من استخدام اللّفظ المُناسب مع الكلمة الصّحيحة.

4. تُعتبرُ اللّغة الإنجليزيّة من أقلِّ اللّغات الأوروبيّة تصريفاً للمفردات والتّراكيب، ويشارُ للتّصريف بأنّه تغييرٌ في شكل أو مبنى الكلمة من أجل المُساهمة في تقدميها لمعنى مُختلِف، أو دور جديد في الجُملةِ.

5. غالباً يأتي الفاعل في الجُملة اللغويّة الإنجليزيّة في بداية الجُملة بعكس الكثير من اللّغات الأُخرى، وخصوصاً اللّغة العربيّة التي تُبدأُ فيها الجُملة الاساسيّة بالفعلِ ومن ثم الفاعل.

# مقارنة بين اللغتين العربية و الانكليزية:

وهنا العديد الاختلافات بين اللغة العربية واللغة الإنجليزية من حيث قواعد اللغة وتراكيب الجمل وترتيب هذه التراكيب و**من أهم هذه الاختلافات ما يلي**:

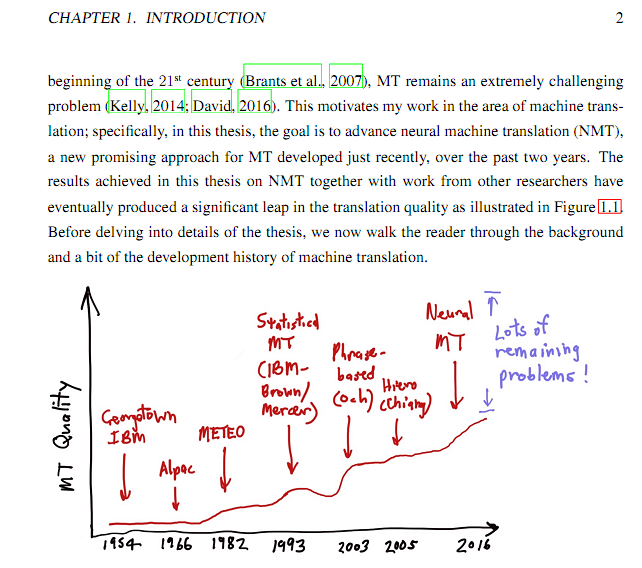
1. ترتيب الفاعل والمفعول به في اللغة العربية تتكون الجملة الفعلية من فعل ويليه الفاعل ثم المفعول به أما في اللغة الإنجليزية فيأتي الفاعل ثم الفعل ثم المفعول به.
2. الجمع والإفراد في اللغة العربية هنا ثلاث أنواع للاسم فإما أن يكون مفردًا أو مثنىً أو جمعًا، أما بالنسبة للغة الإنجليزية فإما أن يكون الاسم مفردًا وإما أن يكون جمعًا.
3. ترتيب الصفة والموصوف في اللغة العربية يأتي الموصوف ثم تتبعه صفته أما في اللغة الإنجليزية فتأتي الصفات التي تتعلق بالموصوف ثم يأتي الموصوف في النهاية.
4. طبيعة الفعل مع المذكر والمؤنث في اللغة العربية يتغير الفعل بحسب جنس الفاعل بحيث يختلف شكل الفعل إذا كان الفاعل مذكرًا عن شكله في حال كان الفاعل مؤنثًا، أما في اللغة الإنجليزية فلا يوجد فرق في شكل الفعل باختلاف جنس الفاعل.

الترجمة الآلية

تعد اللغات البشرية من أهم وسائل الاتصال بين البشر، ومع وجود العديد من اللغات البشرية الموجودة الآن، ومع تطور الحضارة الإنسانية ظهرت الحاجة إلى طريقة اتصال سهلة ومفهومة عابرة للغات البشرية أكثر الحاحاً وبالتالي ظهرت الترجمة الآلية Machine Translation (MT) كمحاولة لسد هذه الحاجة.

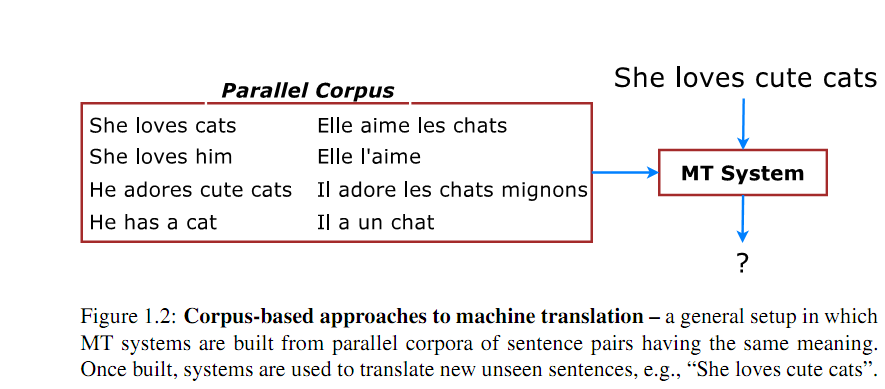
يمكن تعريف الترجمة الآلية هو عملية تدريب الحاسوب لتعلم الترجمة بين اللغات ، الترجمة الآلية لديها تاريخ قديم(Hutchins 2007) حيث وجدت فكرة فلسفية من القرن السابع عشر قائمة على وجود لغات عامة و بدأ التطبيق العملي لهذه الفكرة في بدايات خمسينيات القرن الماضي حيث ظهرت الأفكار الأولى للترجمة الآلية من قبل (Weaver 1949/1955) و التي تعد بداية أبحاث الترجمة الآلية حيث طرح Weaver فكرة استعمال الحواسيب للترجمة، و بعدها مرت الترجمة الآلية بفترات من التطور كما أن أبحاث الترجمة الآلية مرة بفترات من الركود. ظهرت عدة قفزات في هذه المجال (Sheridan 1955).

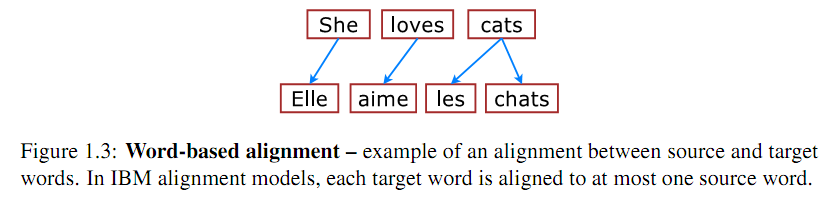
وبعد ذلك بقيت الترجمة الآلية مهمة صعبة، في مطلع القرن الحادي و العشرين ومع تطور أبحاث الشبكات العصبية في مجال التعليم العميق Deeplearnig بدأ الاتجاه باستخدام الشبكات العصبية بأنظمة الترجمة الآلية، و ظهور أنظمة الرجمة الألية بالاعتماد على الشبكات العصبية حيث Neural Machine Translation (NMT) كان هذا قفزة كبيرة في أنظمة الترجمة الآلية.



# تاريخ الترجمة الآلية

كانت بداية أبحاث الترجمة الآلية بين عام 1950-1960 وكانت ترتكز في الغالب على استبدال المباشر للكلمات بناءً على قواميس ثنائية اللغة حيث كان هنالك عمل كبير في هذا المجال والعديد من المقاربات تم بناء الاعتماد على القواميس ثنائية اللغة. و في عام الـ1966 أصدرت اللجنة الاستشارية للمعالجة الآلية للغات The Automatic Language Processing Advisory Committee حيث أشار إنه لا فائدة من أنظمة الترجمة الآلية أوصت بإنشاء أنظمة مساعدة للترجمة البشرية حيث كان لهذا التقرير أثر سلبي في أبحاث الترجمة الآلية و بعدها حدث تباطء كبير في أنظمة الترجمة الآلية لأكثر من عقد، و بالرغم من خفض تمويل أبحاث الترجمة الآلية في الولايات المتحدة، فإن الأبحاث استمرت في أوربا و البلدان الأخرى مع الدعم المقدم من الحكومات و الشركات التجارية ففي عام 1968 أصدرت شركة Systran نظام للترجمة الآلية من اللغة الروسية إلى الإنكليزية حيث استعمل هذا النظام من سلاح الجو الأمريكي في عام 1970 , وفي عام 1976 طورت جامعة مونتريال نظام للترجمة الآلية خاص بتوقعات الأرصاد الجوية (Sloculn 1985)، و بعدها بدأ العديد من أبحاث الترجمة الآلية المعتمدة على القواعد rule-based مثل نظام CARALYSTالذي طور في جامعة Carnegie Mellon نظام Pangloss و الذي طور في جامعة Mexico state و العديد من الأنظمة التي لا يتسع المجال لذكرها.

وبعدها بدأت أبحاث الترجمة الآلية بالتسارع تدريجياُ و بدأت أنظمة الترجمة الآلية الإحصائية بالظهور و خاصة بعد أعمال الباحثين في شركة IBM (Brown، Pietra، وآخرون 1993)حيث أن هذه الأنظمة تعتمد على مجموعة من البيانات على شكل أزواج من الجملة و ترجمتها من لغة إلى أخرى تستخدم لتدريب نموذج للترجمة اللغوية هذه الأنظمة موضحة بالشكل 1.2 و بشكل مفصل أكثر بدلاً من بناء قواميس اللغة بشكل يدوي و التي يمكن أن تكون مكلفة لبنائها ، تم اقتراح تعلم هذه القواميس أو هذه نماذج الترجمة هذه بطريقة احتمالية من المدونات المتوازية , و لتنفيذ هذا الهدف تم اقتراح سلسلة من خمس خوارزميات ليتعلم تقابل الكلمات في المدونات ثنائية اللغة ، و غالباً ما يتم الإشارة إلى النظام بنموذج IBM1-5 و ذلك و هذا ما هو موضح بالشكل1.3

و على الرغم من الأثر الكبير لنماذج IBM في مجال الترجمة الآلية الإحصائية ، سرعان ما اكتشف الباحثون أن الترجمة الآلية القائمة على الكلمات غير كافية و أن الكلمات بحاجة إلى سياق للترجمة بشكل جيد حيث أن الكلمة الواحدة يمكن أن تحتمل أكثر من معنى و بالنتيجة بدأ العمل على نماذج الترجمة الإحصائية بالاعتماد على العبارات phrase-based model (Marcu و Wong 2002، Koehn، Och و Marcu 2003)و أصبحت الطريقة المهيمنة في أنظمة الترجمة الآلية المعيارية حتى وقت قريب حيث يعود الفضل لجهود عمل Och لما يسمى نماذج المقابلة ALIGMENT Template

وفي بدايات العقد الأول من اقرن الحالي بدأ ظهور تطبيقات الترجمة الآلية الإحصائية التجارية google translator Systran ,bing translator , و في مجال اللغة العربية ظهر نظام من أنتاج شركة صخر.

ولا بد من الإشارة أن أنظمة الترجمة الآلية الإحصائية تتألف من العديد من المكونات حيث هناللك العديد من المكونات التي تستخرج المعرفة من النصوص و التي تكون مفيدة في عمليات الترجمة ومن هذه المكونات على سبيل المثال ( النماذج اللغوية ،نموذج الترجمة ، نموذج التدقيق في الترجمة , نموذج التدريب، نموذج الكلمات المجهولة ، طول الجملة ) ولكي يتم من كل هذه الصفات في عملية الترجمة يتم استعمال مقاربة مشهورة في أنظمة معالجة اللغات الطبيعية تدعى maximum-entropy or log-linear model (Berger، Pietra و Pietra 1996) (Och و Ney 2002)

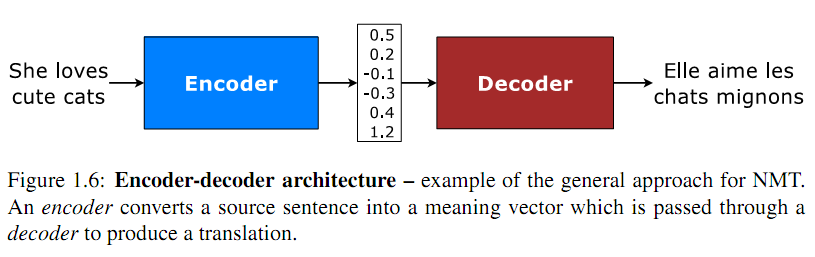
# الترجمة الآلية باستخدام الشبكات العصبية

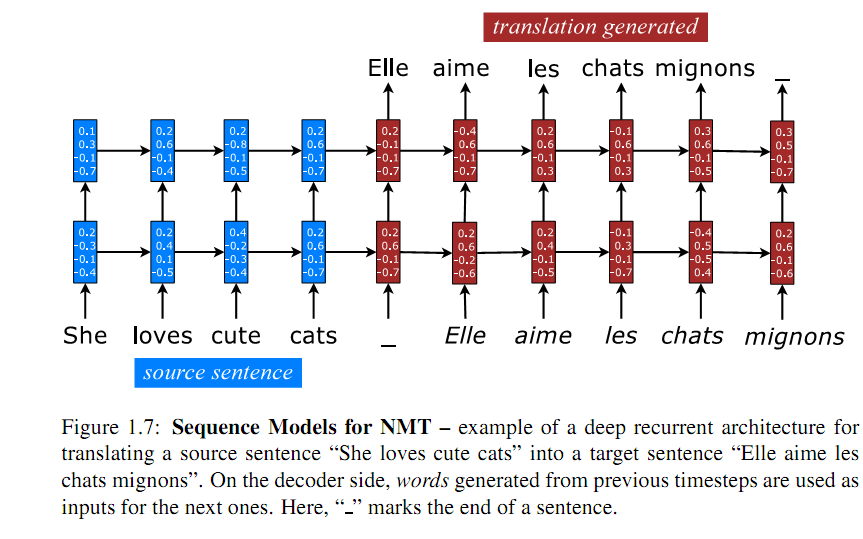
على الرغم من أن الترجمة الآلية الإحصائية قد لاقت نجاحا في العديد من التطبيقات التجارية , فإنها لم تعمل بشكل جيد و عانت من العديد من المشاكل, المشكلة الأولى الترجمة كانت تعتمد على السياق المحلي حيث كانت الترجمة تتمم على مستوى مقطع مقطع و العلاقات بين أجزاء النص البعيدة عن بعضها غالبا ما يتم تجاهلها , المشكلة الأخرى التي عانت منها أنظمة الترجمة الآلية الإحصائية هي أن أجزاء النظام المختلفة معقدة جداً و خاصة عند إضافة خصائص إضافية كما في أنظمة الترجمة الآلية الإحصائية الحديثة (Galley و Manning 2008) (Green, et al. 2013)حيث أن العديد من مكونات النظام بحاجة إلى ضبطها بكل منفصل (أدوات الترجمة ، أدوات الترتيب ، النماذج اللغوية ....) و التي كان من الصعب ضبطها مع بعض ككل و بالتالي إن الترجمة الآلية الإحصائية كانت تعأني و كان ظهور الترجمة الآلية باستخدام الشبكات العصبية هو حل لتلك المشاكل السابق ذكرها (Kalchbrenner و Blunsom، Recurrent continuous translation models 2013) (Sutskever، Vinyals و Le، Sequence to sequence learning with neural networks 2014) (Cho، وآخرون 2014).

في الترجمة الآلية باستخدام الشبكات العصبية(يسمى اختصارأ الترجمة الآلية العصبية)يتألف النظام من شبكة عصبونية واحدة كبيرة جداُ (يمكن أن تصل إلى ملايين العصبونات ) و التي صممت من أجل نمذجة عملية الترجمة بالكامل و كما الحال في أنظمة الترجمة الإحصائية فإن الترجمة الآلية العصبية تتطلب مدونات ثنائية اللغى لكي يتم عملية التدريب ، غير أنها تحتاج إلى معالجة أقل قبل الترجمة، و لكن الخاصية الأهم في الترجمة الآلية العصبية أنها قادرة على التعلم ككل بدون الحاجة لتدريب كل جزء من النظام منفرداً، و بعكس الترجمة الآلية الإحصائية فإن الترجمة الآلية العصبية لها مفهوم سهل نسبيياُ و هو يعتمد على وجود encoder مرمز يقرأ من الجملة الهدف و يقوم بتحيلها إلى شعاع و الذي هو عبارة عن سلسلة من الأرقام هذا الشعاع يقوم بتمثيل معنى الجملة و بعدها يقوم decoder فاك للترميز بمعالجة هذا الشعاع و يقم بتوليد ترجمة كما هو موضح بالشكل 1.7 و غالباً ماتسمى هذه البنية بنموذج encoder-decoder و بهذا تحل الترجمة الآ لية العصبية مشكلة الترجمة المحلية في الموجودة في الترجمة الإحصائية حيث تجمع الترجمة العصبية معلومات عن كل الجملة قبل ترجمتها و بالتالي يمكنها أن تحل مشكلة العلاقات طويلة المدى في اللغات مثل التوافق في الجنس الترتيب في بنية الجملة (فعل فاعل ....) الجمع و الإفراد ...، الشكل 1.6 يوضح بنية العامة لأنظمة الترجمة العصبية

حيث أن الترجمة العصبية تعتمد على الشبكات العصبونية التكرارية Recurrent Neural Networks RNN لكل من المرمز و فاك الترميز (Sutskever، Vinyals و Le، Sequence to sequence learning with neural networks 2014) (Cho، وآخرون 2014).

RNN تكمننا من بناء تمثيل لجملة الدخل ذات الطول المتغير في أنظمة الترجمة الألية العصبية تتكون الشبكة العصبية من شبكتين RNN كما هو موضح بالشكل 1.7، حيث يقوم المرمز أولاً بإنشاء تمثيل للجملة المصدر، وبعدها يقوم فاك الترميز بتوليد جملة الترجمة (الهدف) كلمة في كل مرة حتى يتم توليد رمز خاص يعبر عن نهاية الجملة.





الترجمة الآلية العصبية لديها عدة فوائد إضافية سهولة توليد الجمل ذات الاحتمال الأكبر في مرحلة فك الترميز مقارنة في الخرج المعقد في أنظمة الترجمة الآلية الإحصائية (Koehn، Och و Marcu 2003).

كما أن الترجمة الآلية العصبية يمكن تعميمها بسهولة أكثر بدون أن يكون هنالك تخزين مخصص في جداول الضخمة للمقاطع اللغوية أو النماذج اللغوية.

# الترجمة الآلية واللغة العربية

يمكن اعتبار اللغة العربية من أحد اللغات المهمة التي خضعت لتجارب الترجمة الآلية, و التي بدأت مع الأيام الأولى للترجمة الآلية , خاصة في معاهد و جامعات الولايات المتحدة، و خاصة إذا أخذنا بعين الاعتبار أن الهدف الرئيسي من أبحاث الترجمة الآلية كان من أجل عمليات الاستخبارات و التجسس العسكري و العلمي حيث أخذت أبحاث الترجمة الآلية للغة العربية درجة كبيرة من الأهمية إلى جانب أبحاث الترجمة للغة الروسية في خمسينيات القرن الماضي في الولايات (Vasconcellos 2000) حيث نالت أبحاث الترجمة الآلية دعما من وزارة الدفاع الأمريكية و كانت أبحاث الترجمة الآلية بالنسبة اللغة العربية قد قامت بها Nancy Kennedy.

وعلى الرغم من أن عدة جوانب من اللغة العربية قد تم دراستها في الأيام الأولى للترجمة الآلية، لطالما اعتبرت اللغة العربية وبسبب خصائها اللغوية والصوتية والكتابية من أصعب اللغات المكتوبة والمسموعة لمعالجتها (Boualem 2003).

بدأت الأبحاث الأولى على اللغة العربية في بدايات سبعينيات القرن الماضي ،حيث ركزت الأبحاث الأولى للغة العربية على التحليل المعجمي و الصرفي (lexicons and morphology) و في منتصف التسعينيات ومع اتساع شبكة الانترنت وظهور الحاجة إلى نقل محتويات الويب إلى لغات أخرى (internationalization) و تكاثر وسائل وسائط الاتصال ظهرت الحاجة على بعض من تطبيقات معالجة اللغة العربية و بالتالي بدأت الأبحاث بالاهتمام على جوانب أخرى من اللغة العربية مثل التحليل القواعدي و الترجمة الآلية فهرست الوثائق و استرجاع المعلومات .., التعرف على الكلام ، حيث أظهرت بعض هذه المجالات تقدماً كبيراً.

و على الرغم من وجود عدة مقاربات للترجمة الآلية في اللغات الأوربية إلا أن هذه المقاربات لا يمكن تطبيقها مباشرة للغة العربية و كانت نتائج التطبيق غير مرضية تماما حيث ظهر الاتجاه إلى تبني مساعد للترجمة translation human aided machine (Al\_Wasiti2005)بدلا من أنظمة الترجمة الآلية.

وبدأ في العقد الأول من القرن الحالي ظهور بعض أنظمة الترجمة الآلية للغة العربية كنظام الترجمة الخاص بشركة صخر , بالإضافة إلى شركة Systran و التي أنتجت عدة أنظمة للترجمة الآلية لعدة لغات بما فيها اللغة العربية ، بالإضافة الترجمة الخاص بمحرك البحث Google حيث كانت هذه الأنظمة بمعظمها تعتمد على طرق إحصائية للترجمة و قد خضعت هذه الأنظمة للعديد من الأبحاث لمقارنة جود ترجمة هذه النظم من مختلف الجوانب (Al, Hailat and Al 2013)

ومع بدأ استعمال الشبكات العصبونية في مجال الترجمة الآلية و بدأ ظهور أنظمة تجارية للترجمة بالاعتماد على الشبكات العصبية Google, Microsoft), اتجهت بعض الشركات إلى بناء أنظمة للترجمة العربية بالاعتماد على الشبكات العصبية (Microsoft ,Systran) غير أن هذه الأنظمة في معظمها في مراحل الأبحاث و لم يتم اطلاقها لاستثمار التجاري.

الترجمة الآلية مقدمة نظرية

قبل الخوض في التفاصيل من المفيد أن نعرف بعض المصطلحات التي تظهر في نماذج الترجمة الألية باستخدام الشبكات العصبية و نماذج السلسلة إلى السلسة(Sequence-to-sequence)

ندعو اللغة التي تكون دخل أنظمة الترجمة الآلية باللغة المصدر، ولغة الخرج باللغة الهدف.

وبالتالي يمكن توصيف عمل أنظمة الترجمة الآلية هو تحويل سلسلة من الكلمات من اللغة المصدر إلى اللغة الهدف وهدف الباحث في هذا المجال هو إيجاد نموذج يسمح لنا بهذا التحويل بدقة وفق مجال و اسع من المحتويات.

أما فيما يتعلق بنماذج السلسلة إلى السلسة (Sequence-to-sequence) فهي تشير إلى صنف واسع من النماذج التي تربط سلسلة إلى أخرى و التي أنظمة الترجمة الآلية هي جزء منها و لكن نماذج Sequence-to-Sequence تستعمل في أنظمة أخرى غير أنظمة الترجمة الآلية و التي تعالج مهمات أخرى غير الترجمة الآلية مثل محدد أقسام الكلام التعرف على الكلام أو أية تسلسل من البيانات كما هو موضح بالشكل التالي .

Machine t~~r~~anslation:



kare wa ringo wo tabeta *→* he ate an apple

Tag~~g~~ing:



he ate an apple *→* PRN VBD DET PP

Dialog:



he ate an apple *→* good, he needs to slim down

Speech R~~e~~cognition:



*→* he ate an apple



And just abo~~u~~t anything...:



1010000111101 *→* 00011010001101

Figure 1: An example of sequence-to-sequence modeling tasks.

هنالك حافز كبير بين الباحثين في استعمال أنظمة الترجمة الآلية كممثل عن الطيف الأوسع من أنظمة السلسلة إلى السلسلة وذلك لعدة أسباب

الترجمة الآلية هو مثال جيد و مفيد عن أنظمة السلسلة للسلسة و يمكن أن نستعمل العديد من الأمثلة من الترجمة الآلية التي توضح الصعوبات التي يمكن أن نواجها و كيفية حل هذه المشكلات.

الترجمة الآلية هو المحرك الرئيسي في تطوير أنظمة السلسلة إلى السلسلة والعديد من حلول مشاكل هذه الأنظمة قد تم الوصول إليه عند تطوير أنظمة الترجمة الآلية و بعدها تم تطبيق هذه الحلول في الأنظمة الأخرى.

يوجد بعض الحالات التي استفادت فيه أنظمة الترجمة الآلية من الأنظمة الأخرى والتي ساعدت في حل مشكلات الترجمة الآلية.

# مبادئ الترجمة الإحصائيةStatistical MT Preliminaries

قبل الحديث عن أية نظام محدد للترجمة الآلية ،سنقوم في هذا الفصل بتوصيف آلية عمل الترجمة الآلية الإحصائية كما وضحه (Brown، Pietra، وآخرون 1993)

المطلوب من نمزذج الترجمة الآلية هو ترجمة الجملة المصدر إلى الجملة الهدف

الجملة *التي تبدأ بالكلمة و التي طولها*  ، الجملة *التي تبدأ بالكلمة و التي طولها* .

وبالتالي أية نظام ترجمة ألي يمكن كتمثيله كتابع

|  |  |  |
| --- | --- | --- |
| (1) |  |  |

|  |  |  |
| --- | --- | --- |
| (1) |  |  |

والذي يعطي الفرضية الخرج إذا كان دخله الجملة

أنظمة الترجمة الإحصائية تقوم بالترجمة أولاً عبر أنشاء نموذج احتمالي يعطي احتمالية بالنسبة إلى ) وثانياً إيجاد الجملة الهدف التي تعطي أكبر قيمة لهذا الاحتمال

|  |  |  |
| --- | --- | --- |
| (2) |  |  |

|  |  |  |
| --- | --- | --- |
| (2) |  |  |

حيث هي برامترات النموذج و التي تحدد التوزيع الأحصائي و هذه البرمترات يتم تعلمها من الجمل المترادفة من اللغتين المصدر و الهدف و التي نسميها المدونات المتوازية parallel corpora

وهنا يظهر لدينا ثلاث مشاكل رئيسية و التي يجب أنت نتعامل معها بشكل جيد، لنتمكن من إيجاد نظام جيد للترجمة، وهذه المشاكل هي :

**:Modeling** تحديد ماهية النموذج الذي نريد أن تعامل معه وما هي البارامترات التي يمتلكها، وكيف ستقوم هذه البارامترات بتحديد التوزيع الإحصائي؟

**Learning:** كيفية تعلم القيمة المناسبة لهذه البارامترات من بيانات التدريب.

**Search :** أخيرا التعامل مع مشكلة البحث عن أكثر جملة محتملة (حل معادلة argmax) وعملية البحث عن أفضل فرضية تدعى فك الترميز decoding.

# نماذج n-gram اللغوية

بينما الهدف لأنظمة الترجمة الآلية الإحصائية هو بناء نموذج احتمال جملة الهدف للجملة المصدر , و لكن في هذه القسم سنأخذ خطوة للوراء و سنحاول بناء نموذج للجملة الهدف فقط .

و بشكل رئيسي فإن لهذا النموذج استعمالين رئيسين

1. **تقييم أن الجملة صحيحة:** إذا كان لدينا جملة هل هذه الجملة تبدو جملة فعلية من اللغة الهدف؟ إذا كنا نستطيع تدريب نموذج قادر على فعل هذا فإننا نستطيع استعمال هذا النموذج في تقييم جودة الجمل المولدة من أنظمة الترجمة الآلية وذلك لتحسين جودتها، كما أنها يمكن أن تستعمل لتقييم الجمل المؤلفة من قبل شخص ما للتأكد من القواعد أو تصحيح الأخطاء.
2. **توليد نصوص:** يمكن استعمال هذه النماذج لتوليد نص عشوائي و ذلك عن طريق توليد جملة من التوزيع الإحصائي للغة الهدف هذا النموذج سوف يكون مفيدا أكثر في مجال الترجمة الآلية باستعمال الشبكات العصبية .

## حساب احتمال الكلمات كلمة بكلمة Word-by-word Computation of Probabilities

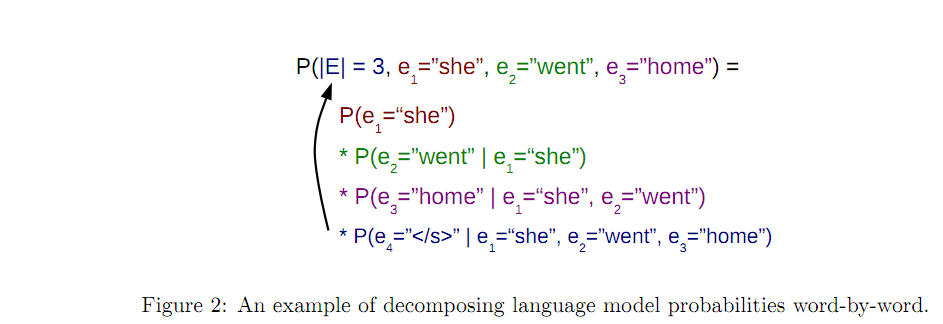
كما ذكرنا سابقاً فإننا مهتمون بحساب احتمالية الجملة و التي يمكن ان نعبر عنها بالعلاقة

|  |  |  |
| --- | --- | --- |
| (3) |  |  |

|  |  |  |
| --- | --- | --- |
| (3) |  |  |

و هو الاحتمال المشترك كون طول الجملة هو و الكلمة الأولى من الجملة و الكلمة الثانية و حتى أخر كلمة في الجملة هي .

الحساب المباشر للتوزيع الاحصائي لهذا النموذج ليس سهل حيث أن طول الجمل ليس محددأ مسبقا و هنالك تركيبات كبيرة جدا للكلمات[[1]](#footnote-2).



هنالك طريقة لتسهيل الأمور وذلك بإعادة كتابة احتمال الجملة كجداء احتمالات كل كلمة على حدا، وبالتالي فإن الاحتمال المشترك يمكن حسابه بجداء الاحتمالات الشرطية لكل كلمة في الجملة.

على سبيل المثال احتمال الجملة يمكن حسابه بالطريقة التالية

الشكل 2 يعطي طريقة لحساب احتمال الجملة "she went home"، بالإضافة إلى الكلمات الفعلية إلى الجملة أضفنا الرمز </s> ليشير إلى نهاية الجملة الموجودة ضمنيا في كل جملة، وبعدها نحسب احتمال went تأتي في المرتبة الثانية في الجمل التي تبدأ بـ she وبعدها حساب احتمال went في الجمل التي تبدأ she went و بعدها احتمال </s> بعد "she went home".

ويمكن أن نعمم ذلك باستخدام العلاقة

|  |  |  |
| --- | --- | --- |
| (4) |  |  |

|  |  |  |
| --- | --- | --- |
| (4) |  |  |

حيث هو <\s> والسبب في إضافة رمز نهاية الجملة هو بأنه يسمح لنا أين أن نحدد لأين تنتهي الجملة أو بعبارة أخرى بفحص مكان <\s> يمكن أن نحدد الحد في المعادلة في مثالنا السابق نعرف أن موضع <\s> هو 4 وبالتالي نعرف أن طول الجملة هو 3.

من المعادلة (3) التي حصلنا عليها أصبحت مشكلة بناء نموذج للغة هي عملية حساب الاحتمال الشرطي للكلمة التالية بالنسبة للكلمات السابقة ، و هذا يمكن معالجته بشكل أسهل من حساب احتمال

كل الجملة حيث لدينا عدد محدد من العناصر التي نريد حساب احتمالاتها،وهذا ما سنبينه في الأقسام

التالية.

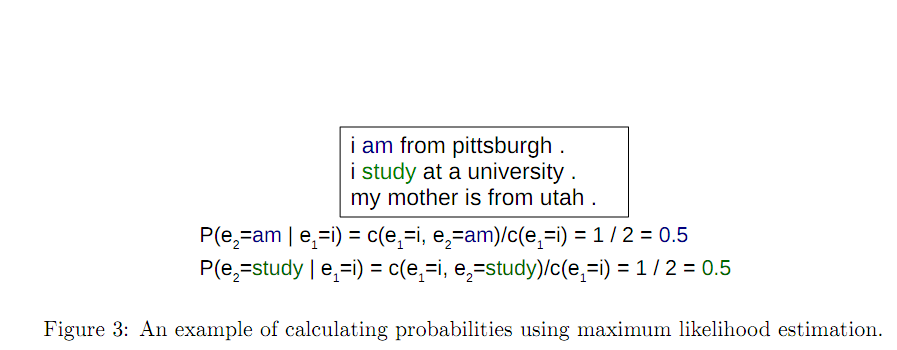
## تماذج اللغة n-gram المعتمدة على العد Count-based n-gram Language Models

الطريق الأولى لحساب الاحتمالات بسيطة : حضر بيانات للتدريب حيث يمكننا منها عد كلمات الجمل ,ومن ثم عد كم مرة شاهدنا فيها كلمة معينة و قسم هذا العدد على عدد المرات التي شاهدنا فيها هذا السياق هذه الطريقة البسيطة يمكن التعبير عنها بالمعادلة التالية

|  |  |  |
| --- | --- | --- |
| (5) |  |  |

|  |  |  |
| --- | --- | --- |
| (5) |  |  |

حيث  *هو عدد المرات التي تظهر فيها كلمة معينة في بداية الجملة في بيانات التدريب.*

*هذه المقاربة تدعى maximum likelihood estimation و هي بسيطة و تضمن أن تعطي احتمالات كبير للجمل التي تظهر في بيانات التدريب .*

ولكن ماذا إذا أردنا أن نحسب احتمال جملة جديدة لم تظهر في بيانات التدريب مثل نريد أن نحسب احتمال الجملة "I am from utah " على بيانات التدريب في المثال السابق هذه الجملة مشابهة للجمل الواردة في بيانات التدريب، و لكن كون أن الجملة " I am from Utah" لم تشاهد في بيانات التدريب فإن

و بالتالي

و بالتالية احتمالية الجملة المحسوبة من المعادلة (5 ستصبع صفر و في الواقع هذا النموذج سيعين احتمالية صفر لكل جملة لم يراها في بيانات التدريب و بالتالي فهو غير مفيد جداً, حيث يفقد النموذج قدرته على إخبارنا أن الجملة الجديدة طبيعية أم لا أو أن يقوم النظام بتوليد جمل جديدة .

لحل هذه المشكلة نقوم بوضع معياريين جديدين، بدلاً من أن نحسب الاحتمالات من بداية الجملة نضع نافذ ثابتة من الكلمات السابقة حيث ستكون قاعدة لحساب احتمالاتنا، ونعتبره تقريباً للاحتمال الحقيقي فإذا حددنا سياقنا بـ n-1 كلمة سابقة فإن هذا سوف يعطينا :

|  |  |  |
| --- | --- | --- |
| (6) |  |  |

|  |  |  |
| --- | --- | --- |
| (6) |  |  |

النماذج التي تعتمد على هذه الفرضية تدعى نماذج n-gram و إذا كانت n=1 فإنها تدعى نماذج unigram ، n=2 تدعى نماذج bigram، n=3 تدعى نماذج trigram،....

بارامترات نماذج n-gramهي الاحتمالات الشرطية للكلمة التالية بالنسبة لـ n-1كلمة سابقة:

|  |  |  |
| --- | --- | --- |
| (7) |  |  |

|  |  |  |
| --- | --- | --- |
| (7) |  |  |

و لكين ندرب نموذج n-gram يجب ان نتعلم هذه البارامترات من بيانات التدريب [[2]](#footnote-3)،هذه البارامترات يمكن حسابها بطريقة maximum likelihood estimation كالتالي

|  |  |  |
| --- | --- | --- |
| (8) |  |  |

|  |  |  |
| --- | --- | --- |
| (8) |  |  |

حيث هو عدد سلاسل الكلمات في أي مكان من بيانات التدريب، بعض هذه المعادلات سوف تشير إلى الكلمات حيث و بالتالي نعتبر  *حيث هو رمز يعبر عن بداية الجملة .*

و إذا عدنا إلى مثالنا السابق و كانت n=2 نلاحظ حتى و إذا لم ترد الجملة "I am from utha" في بيانات التدريب فإن فن السلاسل {"I am","am from","from utah","utha .",". </s>"} قد ظهرت في عينات التدريب و بالتالي نستطيع حساب احتمال غير صفري للجملة من احتمالات السلاسل السابقة.

على كل حال مازال لدينا مشكلة ماذا لو صادفنا تسلسل من كلمتين لم نشاهده في عينات التدريب من قبل سنحصل في هذه الحالة على احتمال صفري من جديد.

n-gram تحل هذه المشكلة و ذلك **بتنعم الاحتماللات** و ذلك بجمع طريقةestimate maximum likelyhood من نماذج مختلفة ذات قيم مختلفة لـn في أبسط الحالات ندمج احتمالات نماذج unigram وbigram،يكمن اعتبار النموذج الذي يجمع الاحتمالات كالتالي:

|  |  |  |
| --- | --- | --- |
| (9) |  |  |

|  |  |  |
| --- | --- | --- |
| (9) |  |  |

حيث هو متحول يحدد كمية الاحتمالات التي نحفظها لنموذج unigram، و طالما و بغض النظر عن السياق كل الكلمات سيكون لديها احتمال ما هذه الطريقة تدعى **interpolation** وتعتبر من الطرق الفعالة في معالجة الظواهر قليلة التكرار.

و إذا أردنا ان نستخدم سياق أطول n=3, n=4,n=5 يمكن بطريقة عودية أن ندمج الاحتماللات كالتالي

|  |  |  |
| --- | --- | --- |
| (10) |  |  |

|  |  |  |
| --- | --- | --- |
| (10) |  |  |

حيث الحد الأول هو maximum likelihood estimate للنموذج من الدرجة m والحد الثاني هو احتمال الناتج عن دمج النماذج حتى m-1.

هنالك عدة طرق أكثر تعقيدا للتنعيم وهي خارج مجال هذا البحث ولكن يمكن تلخيصها كالتالي (S. F. Goodman 1996)

**معاملات التنعيم المعتمد على السياق:** بدلا من ان يكون لدينا ثابت، نولف معامل الدمج بناءاً على السياق هذا يسمح للنموذج أن يعطي أوزان أكثر لنماذج n-gram ذات المرتبة الأعلى عندما يكون هنالك عدد كافة من عينات التدريب، و يعود إلى النماذج ذات المرتبات الأدنى عندما تكون الأمثلة في بيانات التدريب.قليلة معاملات التنعيم المعتمدة على السياق يمكن اختيارها تجريباً (I.H. Witten and T.C. Bell 1991) أو يمكن تعلمه من بيانات التدريب (Dyer 2016).

**Back-off:** في المعادلة رقم (9) ندمج توزيعين احصائيين لكل المفردات ، في صيغة لـBack-off المغايرة، التوزيع الاحتمالي للنماذج ذات الدرجة المنخفضة يتم استعماله حين يتم يعطينا نموذج ذي درجة أعلى احتمالاً صفرياً هذه التقنية أكثر تعبيرا و لكنها أكثر صعوبة, و لكن النتيجتين أعطتا نتائج متقاربة (J. Goodman 2001)

**التوزيع الاحتمالي المعدل:** يمكن استعمال توزيع احصائي غير Maximum likelihood estimation و ذلك بطرح قيمة ثابتة من العدد قبل حساب الاحتمالات و هذه الطريقة تدعى **discounting** .

الطريقة المعتمد حاليا لتنعيم نماذج n-gram هي Modified Kneser-Ney smoothing (Chen and Goodman, An empirical study of smoothing techniques for language modeling 1999) وتعتبر طريقة فعالة لتنعيم نماذج n-gram وهي تستعمل مزيجا من الطرقا السابقة لتعطي تقدير أدق.

**تقييم النماذج اللغوية**

إذا كان لدينا نموذج لغوي وأردنا أن تختبر إذا كان يعمل بشكل مناسب نتبع الأسلوب التقليدي المتبع في نماذج تعليم الآلة و ذلك بتحضير ثلاث مجموعات من البيانات

1. بيانات التدريب: وتستعمل لتدريب بارامترات النموذج
2. بيانات التحقق: تستعمل للاختيار بين النماذج المختلفة أو تولف البارامترات الفوقية (hyper-parameters) في النموذج ويمكن أن تكون القيمة العظمى لـ n في نماذج n-gram أو اختيار طريقة التنعيم المناسبة
3. بيانات الاختبار يستعمل لحساب دقة النموذج النهائية و كتابة تقارير عن النتائج.

للنماذج اللغوية نريد أن تحقق من أن النموذج الناتج هو نموذج دقيق للغة: و هنالك عدة طرق لتحديد هذا، أكثر الطرق وضوحاً لتحديد الدقة هو استعمال احتمالية **likelihood** للنموذج بالنسبة لبيانات التحقق، احتمالية likelihoodالبرمترات بالنسبة إلى هذه البيانات يساوي الاحتمالية التي يعطيها النموذج لهذه البيانات، على سبيل المثال إذا كانت بينات الاختبار فإن هذه الاحتمالية هي

|  |  |  |
| --- | --- | --- |
| (11) |  |  |

|  |  |  |
| --- | --- | --- |
| (11) |  |  |

غالبا مانفترض أن الجمل في بيانات التدريب مستقلة احتماليا وبالتالي:

|  |  |  |
| --- | --- | --- |
| (12) |  |  |

|  |  |  |
| --- | --- | --- |
| (12) |  |  |

معيار شائع أخر و أكثر استعمالاً هو لوغاريتم الاحتمالية log likelihood

|  |  |  |
| --- | --- | --- |
| (13) |  |  |

|  |  |  |
| --- | --- | --- |
| (13) |  |  |

لوغارتيم الاحتمالية يستعمل لسببين رئيسيين

احتمالية أي جملة وفق النموذج ستكون صغيرة جدا و جداء هذه الأرقام الصغيرة جدا يمكن أن تكون صغيرة جدا مما يسبب مشكلة في دقة النتائج، السبب الثاني هو أنه من الأفضل التعامل في المجال اللوغارتمي و خاصة عند اجراء المشتقات في طرقالرياضية لتحسين البرمترات و التي يستم الحديث عنها في الفصول القادمة و انه من الأفضل التعامل مع الجمع في المعادلة 13 بدلاً من الضرب في المعادلة 11.

و من الشائع تقسيم أحيانا تقسيم لوغارتيم الاحتمالية على عدد الكلمات في البيانات.

|  |  |  |
| --- | --- | --- |
| (14) |  |  |

|  |  |  |
| --- | --- | --- |
| (14) |  |  |

وهذا يجعل من السهل النتائج المتباينة من المدونات ذات الاطوال المختلفة

معيار أخر لقياس دقة النماذج اللغوية هي perplexity و التي تعرف بانها تابع اسي لمتوسط ناقص لوغاريتم الاحتمالية لكل كلمة.

|  |  |  |
| --- | --- | --- |
| (15) |  |  |

|  |  |  |
| --- | --- | --- |
| (15) |  |  |

يمكن تفسير الـperplexity بانه مدى ارتباك النموذج من قراره و بشكل دقيق أكثر بأنه يعبر عن قيمة الناتجة عند اختيار كلمات عشوائية من التوزيع الإحصائي الذي تم حسابه من قبل النموذج اللغوي عند كل خطوة زمنية ,وسطياً كم عدد الكلمات التي يجب اختيارها حتى نحصل على الكلمة الصحيحة ؟ و السبب في شيوع هذا المعيار كلما كان الرقم المحسوب منه أكبر كان افهمنا لهذه النماذج أسهل من قبل الإنسان.

## معالجة الكلمات المجهولة

أحد النواحي الهامة التي يجب الانتباه إليها هو أنه بعض الكلمات في مجموعة الاختبار  *لن تظهر و لا لمرة و احدة في مجموعة التدريب هذه الكلمات تدعى الكلمات المجهولة unknown words، و يجب معالجتها بطريقة ما، الطرق الشائعة في معالجة هذه المشكلة هي.*

***فرضية القاموس المغلق*** *والتي تفترض عدم وجود كلمات جديدة في مجموعة التحقق، وبالتالي نعطيها احتمالية صفر و بالتالي يصبح احتمال الجملة المختبرة صفر أيضا هذه الفرضية مفيدة في مجال ضيق من التطبيقات و لا يمكن تعميمها على الحالات الأخرى.*

***دمجها مع التوزيع الإحصائي للكلمات الغير معروفة :****كما هو مذكور في المعادلة* 10 *يمكننا ان ندمج التوزيعات الإحتمالية لنماذج ذات درجة عليا و دنيا و بالتالي يمكن اعتبار الكلمات المجهولة تابعة لتوزيع احصائي لنموذج ذي درجة 0 و نعرف توزيع الإحصائي لنموذج 1-gram هو الدمج بين توزيع نموذج unigram و توزيع نموذج الكلمات المجهولة*

|  |  |  |
| --- | --- | --- |
| (16) |  |  |

حيث يجب ان يكون التوزيع الذي يعطي احتمالة لكل الكلمات  *و ليس فقط للمفردات V الموجودة في قاموسنا المشتق من بيانات التدريب. و هذا يمكن تنفيذه على سبيل المثال بتدريب نموذج لغوي على مستوى الأحرف يقوم بتهجئة الأحرف في الكلمات المجهولة الغير موجودة في مفرداتنا أوبشكل أبسط و افضل من تجاهل الكلمات المجهولة هو تقدير عدد الكلمات الكلي*   *في اللغة التي نريد نمذجتها حيث*  *و نعرف*  كتوزيع منتظم لكل الكلمات ؛

إضافة الرمز <unk> : و يمكن اعتبارها الطريقة الأخيرة لمعالجة الكلمات المجهولة حيث نقوم بإزالة بعض الكلمات من بيانات التدريب و نستبدلها بالرمز <unk> الذي يمثل الكلمات المجهولة و أحد الطرق المشهورة لفعل هذا هو إزالة الكلمات المفردة أو التي تظهر مرة و احدة في بيانات التدريب، و بفعل هذا سنكون بصراحة حددنا السياق الذي سترد فيه الكلمات المجهولة بدل من أن نتوقعها ضمنيا من خلال دمج النماذج كما هو مذكور سابقا، وحتى توقعنا الرمز <unk>مازلنا بحاجة على احتمالية الكلمة الحقيقية وبالتالي كل مرة نتوقع فيها <unk> في الموضع نقوم بالضرب باحتمال .

تحدثنا في هذا القسم عن النماذج اللغوية n-gram و لم نتوسع في الكلام عنها إلا لنعرض المشاكل التي تعانيها هذه النماذج و يمكن لهذه النماذج أن تكون كبيرة جدا و جزء من الأنظمة التجارية بحيث يتم بناء هذه النماذج بالاعتماد على حجم هائل من البيانات تم الحصول عليه من الويب أو مصادر أخرى (Heafield 2011) (Pauls and Klein 2011)،ويمكن أن تتوزع هذه النماذج اللغوية الضخمة على عدة حواسي (Brants, et al. 2007) كما يمكن تخصيص هذه النماذج لتتبنى مجال معين (Talbot and Brants 2008) .

**النماذج المعتمد على العد بعيدة المدى**

كما تم ذكره سابقا نماذج n-gram تحصر سياقها بـ n-1 كلمة سابقة و لكن بالواقع هنالك اعتمادية بين الكلمات أصول من ذلك و يمكن حتى ان تتجاوز الجملة نفسها و يمكن حتى ان تمتد بين الوثائق، نماذج الشبكات العصبية التكرارية التي سنتحدث عنها لاحقاً أحدى الطرق لمعالجة هذه المشكلة و لكن هنالك طرق غير الشبكات العصبية لحل هذه المشكلة كالنماذج اللغوية المعتمدة على ذاكرة التخزين المؤقتة (Kuhn and De Mori 1990) ، نماذج الموضوعات (Blei, Ng and Jordan 2003)،skip-gram models (J. T. Goodman 2001).

**النماذج المعتمدة على القواعد اللغوية:**

و هذ النماذج تأخد في الحسبان قواعد الجملة الهدف، يمكن على سبيل المثال تكييف الاحتمالات على الكلمات القريبة قواعديا بخلااف نماذج n-grma و التي تكون الاحتمالات فيه قائمة على أساس الكلمات القريبة من بعضها قواعدياً (Shen, Xu and Weischedel 2008)

# 4النماذج اللغوية الخطية الوغارتمية[[3]](#footnote-4)

سنناقش في هذا الفصل مجموعة أخرى من النماذج اللغوية و هي التماذج اللغوية الخطية (Rosenfeld 1996) (Chen and Rosenfeld 2000)و التي تأخذ مقاربة أخرى بدلامن نماذج n-gram المعتمدة على العد

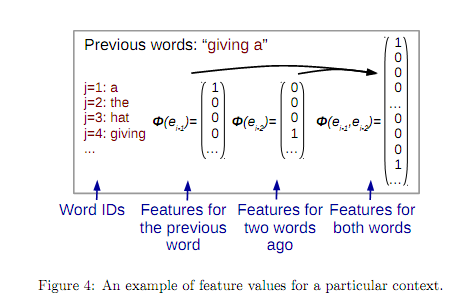
## 4.1صياغة النموذج

كما في نماذج n-gram نماذج اللغوية الخطية اللوغارتمية تحسب أيضاً احتمالية كلمة معينة في سياق معين ،و لكن طرقهم لحساب ذلك مختلفة تماماً و ذلك باتباع الخطوات التالية

**حساب الخصائص:** الانظمة الخطية اللوغاريتمية قائمة على مفهوم **الـخصائص،** و باختصار الخاصية هي "شيءما عن لسياق و الذي سيكون مفيداً في توقع الكلمة التالية"، و بشكل علمي أكثر نعرف تابع الخصاية الذي يأخذ السياق كدخل و يعطينا شعاع من الخصائص و الذي هو عبارة عن شعاع من القيم الحقيقية و الذي يصف السياق بـN خاصية مختلفة.

على سبيل المثال في نماذج bi-gram نعرف أن معرف الكلمة السابقة شيء مفيد لتوقع الكلمة التالية، و إذا أردنا أن نعبر عن الكلمة السابقة بشعاع ذو قيم حقيقية، نعتبر كل كلمة من قاموس مفرداتنا لها معرف خاص j و بعدها نعرف تابع الخاصية ليعيد شعاع الخصائص في هذا الشعاع يكون الموقع رقم j يساوي 1 و بقية العناصر يساوي الصفر ، هذا النوع من الترميز يدعى one-hot vector و هو ما موضح في الشكل رقم 4-a.

ولاحقاً سنعرف التابع و الذي يعطينا شعاع يكون الموقع رقم يساوي 1 و بقية العناصر يساوي الصفر, حيث نفترض أن طول الشعاع مناسب للسياق المعطى.



بالطبع لسنا مجبرين على اعتبار كلمة و احدة سابقة فقط يمكننا حساب أشعة One-hot لكلا و و بعدها نوصلهم مع بعضهم و لنحصل على نموذج يأخذ بعين الاعتبار الكلمتين السابقتين , في الواقع هنالك أنوع أخرى من توابع الخصائص يمكننا استعمالها و التي سنتحدث عنها في الفصل 4\_4, و المرونة التي نملكها في تعريف تابع الخصائص هي ميزة لنماذج log-liner عن نماذج n-gram.

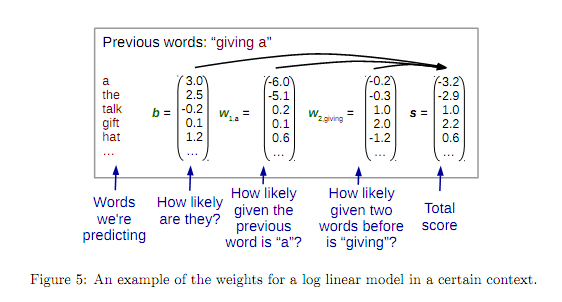
**حساب العلامات:** بعد حساب شعاع الخصائص نريد استعمال هذه الخصائص لحساب احتمالات المفردات ، و لكي نقوم بهذا نقوم بحساب شعاع العلامات و الذي يمثل احتمالية likelihood كل كلمة في المفردات الكلمات ذات العلامات العالية ستكون ذات احتمال أكبر. حيث سنقوم بذلك باستعمال بارامترات النموذج و التي تأتي في نوعين **شعاع الانحيازات** و التي تخبرنا عن قوة كلمة في المفردات و مصفوفة الأوزان و التي تعطينا العلاقة بين الخصائص و الأوزان , و هكذا المعادلة الأخيرة لحساب النتائج هي :

|  |  |  |
| --- | --- | --- |
| (17) |  |  |

شيء آخر يجب ملاحظته في بعض الحاللات الخاصة في الأشعة one-hot أو الأنواع الأخرى من الأشعة و التي تكون معظم عناصرها صفر و و بالتالي يمكن إعادة صياغة المعادلة 17 بطريقة أخرى رياضية ولكنها أثر فعالية حسابيا و بدلاأ من ضرب شعاع ذو عدد كبير من العناصر بمصفوفة الأوزان الكبيرة الأبعاد يمكت جمع أعمدة المصفوفة لكل الخصائص الغير صفرية.

|  |  |  |
| --- | --- | --- |
| (18) |  |  |

حيث هو العمود من مصفوفة الأوزان ، يمكن التفكير بالعملية الحسابية كالبحث عن الخصائص الفعالة و جمعهم مع بعضهم، المثال التالي يوضح هذه العملية حيث عدد الخصائص 2 خاصية للكلمة السابقة و خاصية لكلمة قبل السابقة.



**حساب الاحتمالات:** يجب الملاحظة ان العلامات s هي أرقام حقيقية و ليست احتمالات و يمكن أن تكون سالبة أو أكبر من 1 , و لا يوجد قيود بحيث يكون مجموعها 1 ، ومن أجل هذا نمرر هذا العلامات على تابع يجري التحويل التالي:

|  |  |  |
| --- | --- | --- |
| (19) |  |  |

بأخذ التابع الأسي للعلامات و تقسيمها على مجموع التابع الأسي للعلامات على كل المفردات نحول هذه القيم إلى احتماللات قيمها بين 0 و 1 و مجموعها 1.

هذا التابع يدعى softmax و غالباً ما يتم التعبير عنه بالشكل الشعاعي

|  |  |  |
| --- | --- | --- |
| (20) |  |  |

و من خلال تطبيق هذا التابع على العلامات التي تم حسابها في الفقرة السابقة نكون قد انتقلنا من الخصائص إلى احتمالات النموذج اللغوي .

## 4.2تعلم البرمترات

بقي لدينا كيف سنقوم بتعليم برمترات النموذج ,الأوزان W و الانحيازات b، نقوم بهذه العملية عن طريق إيجاد البرمترات التي تناسب بيانات التدريب جديداً.

للقيام بهذا نستعمل طرق تعليم الآلة لأيجاد القيمة المثلى للبرمترات أول نعرف **تابع الخطاُ loss function** و الذي يبين مدى سوء النتائج التي يعطيها النموذج على بيانات التدريب و في هذه الحالة نعتبر هذا التابع هو **ناقص لوغارتيم الاحتمالية** أو ما يسمى **crossentropy loss:**

|  |  |  |
| --- | --- | --- |
| (21) |  |  |

نفترض يمكن ان نعرف الخطأ علة مستوى الكلمة كالتالي

|  |  |  |
| --- | --- | --- |
| (22) |  |  |

الخطوة التالية هي تحسين البارامترات لتقليل هذا الخطأ، و بينما هنالك العديد من الطرق لإجراء هذه المهمة، أحد أهم هذه الطرق و أسهلها هي stochastic gradient descent(SGD).

SGD عملية تكرارية حيث نقوم باختيار كلمة عشوائية و نقوم بخطوة لتحسين الاحتمالية بالنسبة لـ حيث نقوم بحساب مشتقات تابع الخطأ بالنسبة لكل برامتر من برمترات النموذج .

|  |  |  |
| --- | --- | --- |
| (23) |  |  |

وبعدها نستخدم هذه المشتقات لتعديل قيم بارامترات النموذج وذلك بالقيام بخطوة بحيث يتم تقليل تابع الخطأ وفق العلاقة التالية:

|  |  |  |
| --- | --- | --- |
| (24) |  |  |

حيث هو **معدل التعلم** و وهو يحدد مدى التعديلات التي نجريها على البارامترات في كل مرة نقوم بتعديلها, حيث تعديل الخطوات يؤدي إلى تقليل الخطأ او زيادة احتمالية بيانات التدريب.

هذه الصيغة من SGD بسيطة و فعالة جدا لتحسين الأنظمة الكبيرة و حتى الضخمة منها، و لكن هنالك بعض الأمور التي يجب أن نأخذه بعين الحسبان للحصول على أفضل النتائج.

**اختيار معدل التعلم :** حيث اختيار معدل التعلم بشكل جيد حيث إذا كانت قريبة يمكن أن يصبح التدريب غير مستقر بل حتى يمكن أن تتزايد قيمة تابع الخطأ، , إذا اخترناها صغيرة جداً يمكن أن يصبح تدريب النموذج بطيء جداً و يمكن الوصول إلى نهاية محلية صغرى، أحد الطرق لحل هذه المشكلة هو **تخفيض معدل التعلم** **Learning rate decay** حيث تبدأ بمعدل تعلم كبير نسبياً و بعدها ننقص معدل التعلم مع نهاية التدريب.

**توقيف التدريب المبكر:** كما ذكرنا سابقاً عند تدريب النموذج فإنه يوجد لدينا مجموعتين من البيانات بيانات التدريب وبيانات التحقق حيث أن تدريب الأوزان فقط يتم على بيانات التدريب أم بيانات التحقق فهي لاختبار خطأ النموذج على هذه البيانات حيث نقوم بحفظ النموذج الذي يعطي أقل قيمة خطأ على بيانات التحقق. حيث أن النموذج يبدأ بـoverfiiting (الملائمة الزائدة لبيانات التدريب و بالتالي يقد القدرة على التعامل بشكل جيد مع البيانات الجديدة generalize و أحد الطرق لحل هذه المشكلة هو التوقف عن التدريب عندا يبدأ خطأ النموذج على بيانات التحقق يتوقف عن التحسن أو يبدأ بالتزايد، إجراء آخر يمكن تطبيقه هو تخفيض معدل التعلم.

**إعادة ترتيب البيانات:** أحد خصائص SGD بأنها تعالج بيانات التدريب كتلة و احدة كل مرة .و رغم ان هذه الطريقة جيدة و فعالة, و لكن هذا يمكن أن يسبب مشكلة إذا كان هنالك تحيز في ترتيب بيانات التدريب، على سبيل المثال إذا كانت بيانات التدريب عبارة نصوص نشرات الأخبار حيث تبدأ بالأخبار السياسية ثم الرياضية ثم الترفيهية فإنه من المحتمل أن تكون الأمثلة الأخيرة في بيانات التدريب عبارة عن المئات بل الآلاف من الأخبار الترفيهية و بالتالي فإن بارامترات النموذج ستتعدل في الاتجاه المناسب للأمثلة التي شاهدها حديثاً و لحل هذه المشكلة من المفضل و أعادة ترتيب هذه البيانات بشكل عشوائي لكسر أي ارتباط ضمني موجود بين أمثلة التدريب.

وثمة ايضا عدد اخر من القواعد المستكملة التي اقترحت لتحسين التدرج وجعله أكثر استقرارا او كفاءة. وهذه الطرق هي:

**SGD with momentum** (Rumelhart, Hinton and Williams, Explorations in the microstructure of cognition 1986): بدلاً من أخذ خطوة واحدة في اتجاه التدرج الحالي SGD with momentum تأخذ في الحسبان التدرجات السابقة بشكل متخافت، و هذا يجعل التعلم أكثر نعومة في فضاء البارامترات.

**ADdGrad** (Duchi, Hazan and Singer 2011)**:** هذه الطريقة تعتمد على أن بعض البارامترات يتم تعديلها بشكل أكثر من غيرها، في بعض الحالات يتم تعديل الانحيازات بشكل دائم ، وبعض الأوزان نادرا ما يتم تغيرها ،لذلك فإن ADdGrad تعدل عامل التعلم لكل بارامتر على حدا، بحيث أن البارامترات التي يتم تعديلها كثيرا تحصل على تعديلات أصغر، و البارامترات التي نادرا ما يتم تعديلها تحصل على تعديلات أكبر.

(Kingma and Ba 2014)**Adam:** وهي طريقة أخرى لحساب معدل التعلم لكل بارامتر على حدا، و ذلك عن طريق أخذ متوسطات و تباينات التدرجات السابقة، وذلك بدمج طريقتي momentum و AdaGrad مع بعضهما، وهي أكثر الطرق الشائع لتحسين البارامترات، حيث تزيد سرعة تدريب النموذج و تسهل اجراء العديد من التجارب , و لكنها معرضة أكثر للمشكلة الـoverfitting و بالتالي يجب استعمالها بحذر و مقارنتها بـSGD المعيارية.

## 4.3 حساب مشتقات النماذج اللوغاريتمية الخطية

العملية الأهم عند حساب الأنظمة اللوغاريتمية الخطية هو حساب مشتقات تابع الخطأ بالنسبة لبارامترات وللقيام بهذا نقوم بحساب تابع الخطأ بمرور واحد وفق العلاقات:

|  |  |  |
| --- | --- | --- |
| (25) |  |  |
| (26) |  |  |
| (27) |  |  |
| (28) |  |  |

و باستخد قاعدة chain rule في حساب المشتقات

|  |  |  |
| --- | --- | --- |
| (29) |  |  |
| (30) |  |  |

و بتطبيق العلاقات السابق نجد مشتق تابع الخطأ بالنسبة للانحيازات و لكل عمود في مصفوفة الأوزان هي :

|  |  |  |
| --- | --- | --- |
| (31) |  |  |
| (32) |  |  |

## 4.4 خصائص أخرى للنماذج اللغوية

أحد الأسباب في كون النماذج اللغوية الخطية جيدة هو المرونة التي تعطينا يها عند اختيار الخصائص التي نعتقد أنها مفيدة عند توقع الكلمة التالية و بعض هذه الخصائص :

**كلمات السياق:** كما ذكرنا في المثال السابق يمكن اختيار الكلمة أو الكلمة .

**صنف السياق:** كلمات الساق يمكن أن يتم تجميعها إلى أصناف من الكلمات المتشابهة كطريقة Brwon للتجميع cluster (Brown, Desouza, et al. 1992)، و بدلاً من البحث في شعاع one-hot حيث يمثل كل بعد كلمة واحدة ،يمكن أن نبحث في شعاع one-hot حيث كل بعد يمثل صنف واحد (Chen, Shrinking exponential language models 2009). و بالتالي الكلمات من الصنف الواحد تتشارك نفس الثقة الإحصائية. و بالتالي تمكن النموذج من أن يعمم بشكل أفضل.

**السياق المعتمد على الإضافات suffixes :** و هنا ننظر إلى اللواحق المضافة للكلمة مثل (ing) و بالتالي يتعلم النموذج بشكل أفضل عن الأفعال المستمرة .

**حقيبة الكلمات Bag-of-words:** بدلاً من النظر n كلمة سابقة يمكن استعمال كل الكلمات السابقة في الجملة ,و باستعمال شعاع one-hot لكل كلمة، لكن عوضاً عن وصل الأشعة مع بعضها نقوم بجمعها مع بعضها, لكن هذا سيفقدنا معلومات عن موضع كل كلمة، ولكن نستطيع الحصول على معلومات عن الكلمات التي تأتي مع بعضها في الجملة.

# الشبكات العصبية التكرارية Recurrent Neural Network RNN

## 5.1مقدمة

الشبكات العصبية التكرارية هي أداة فعالة للتعامل مع البيانات المتسلسلة (Elman 1990)وقد تم تطبيقها في مجال النماذج اللغوية (Mikolov, Martin, et al. 2010) (Mikolov, Stefan, et al. 2011) (Mikolov and Zweig 2012)

حيث تأخذ شبكات RNN الدخل كسلسة أشعة و تعالجه واحد بعد واحد حيث لكل شعاع دخل جديد ، تعدل شبكة RNN ذاكرتها الداخلية لتوليد حالة داخلية و التي يمكن اعتبارها كتمثيل للسلسة الجزئية فعالية شبكات RNN تكمن في كيفية تعديل حالتها الداخلية ,حيث يمكن اعتبار الحالة الداخلية في الخطوة t هو تابع للدخل في الخطوة t و الحالة الداخلية في اللحظة t-1

|  |  |  |
| --- | --- | --- |
| (33) |  |  |

الحالة الداخلية البدائية غالباُ ما يتم أسناده إلى الصفر أو أية قيمة أخرى كما سنرى لاحقاً، من الشائع أن يكون f هو تابع غير خطي كـsigmoid أو tanh .

|  |  |  |
| --- | --- | --- |
| (34) |  |  |

عند كل خطوة زمنية يمكن لشبكة RNN تعطي خرج (اختياري) و الذي بمكن أن يكون متقطع أو قيمة حقيقية و في حال كان الخرج متقطع و هذه هي حالة النماذج اللغوية يتم حساب توزيع احتمالي لأصناف الخرج وفق العلاقات التالية:

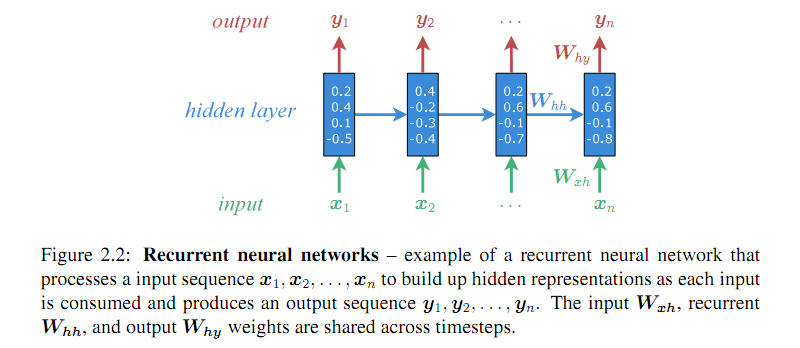
|  |  |  |
| --- | --- | --- |
| (35) |  |  |
| (36) |  |  | |

حيث هي مصفوفة الأوزان بين الوحدات الخفية و الوحدات الخرج و هو حجم الحالة الخفية، و في حال كان لدينا مجموعة أصناف كبيرة جدا للخرج يصبح الجداء بين المصفوفة و الشعاع في المعادلة 35 مشكلة حسابية في شبكات RNN و التي نستعمل في النماذج العصبية اللغوية و الترجمة الآلية سنناقش هذا المشكلة في الفصول القادمة.

حيث تابع softmax تم ذكره سابقاً يحول شعاع العلامات إلى شعاع من الاحتمالات حيث يحدد احتمال كل صنف من أصناف الخرج و يحسب من العلاقة 19.

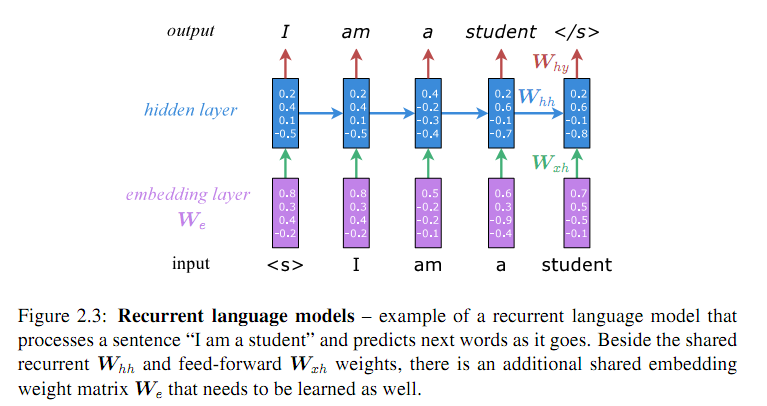
تنقسم أوزان شبكات RNN إلى ثلاث مجموعات ،الأوزان بين طبقة الدخل و الطبقة الخفية ،و الأوزان في الطبقة الخفية أو الأوزان الدورانية , أخيراُ الأوزان بين الطبقة الخفية و الخرج .

هذه الأوزان يتم مشاركتها بين مختلف الخطوات الزمنية كما هو موضح بالشلكل 2.2 ،و هذه الخاصية هي سبب فعالية شبكات RNN، إذ انها يمكن أن تنمذج سلاسل ذات أطول عشوائية دون أن يكون هنالك حاجة لزيادة سعتها، بخلاف شبكات feedforward التي يمكن أن تحدد العلاقات فقط في المقاطع ذات الثابت.

وعلى الرغم من أنه لشبكات RNN العديد من التطبيقات إلا أنه لن نتحدث عنها إلا في مجال تطبيقات النماذج اللغوية.

## 5.2النماذج اللغوية بالاعتماد على شبكات RNN

كحالة خاصة من شبكات RNN فإن التماذج اللغوية المعتمدة على شبكات RNN نفترض أن سلاسل الدخل و الخرج مؤلفة من عدد محدد من الرموز و غالبا ما تكون هذه الرموز هي كلمات اللغة. و غالباً ما يضاف للسسلة محرف خاص يعبر عن بداية السلسلة مثل، و بما أنه الهدف من النماذج اللغوية هو توقع الكلمة التالية ، فإن سلسلة الخرج مزاحة بمقدار 1 عن من سلسلة الدخل كما أنها تنتهي بالرمز </s> و الذي يدل على نهاية السلسلة حيث يصبح المثال السابق كما هو موضح بالشكل 2.3 حيث الكلمة الناتجة في أحد الخطوات الزمنية تصبح كدخل في الخطوة الزمنية التالية .



لتطبيق شبكات RNN على جمل اللغات الطبيعية أو بشكل عام على تسلسل من الرموز المتقطعة يمكن الاعتماد على ترميز one-hot للكلمات حيث حيث هو قائمة المفردات لمستخدمة، و لكن إذا كانت قائمة المفردات كبيرةيصبح هذا التمثيل غير جيد على الإطلاق، حيث سيؤدي على ان مصفوفة الأوزان سوف تكون كبيرة جداً، بالإضافة أنه ليس هنالك أية علاقة تسمح لنا بإيجاد درجة القرابة بين كلمتين مختلفتين . وفي الناحية العملية فإنه يستعمل تمثيل منخفض الأبعاد للكلمات او ما يسمى embeddings لحل المشكلتين السابقتين، وبشكل أدق فإن مصفوفة الـ embeddings يتم تدريبها من أجل الحصول على تمثيل لكل كلمة من كلمات المفردات حيث ، و بالنتيجة فإن النماذج اللغوية لشبكات RNN سيكون لها أوزان  *.*

## 5.3 تدريب شبكات RNN بخوارزمية BackPropagation

هدف التعلم في شبكات RNN هو تصغير ناقض لوغاريتم الاحتمالية او ما يسمى crossentropy loss لأمثلة التدريب، إذا كان لدينا N سلسلة خرج من القيم المتقطعة و ذات أطوال و بالتالي فإن تابع الخطأ يمكن التعبير عنه بالعلاقة:

|  |  |  |
| --- | --- | --- |
| (37) |  |  |

|  |  |  |
| --- | --- | --- |
| (38) |  |  |

غالباً مايتم تدريب شبكات RNN باستخدام خوارزمية mini-batch stochastic gradient descent(SGD) ، حيث مجموع صغيرة من بيانات التدريب mini-batch لحساب المشتقات و تعديل الأوزان في كل مرة .

*استعمال mini-batch لديه عدد من المحاسن*

* أكثر وثوقيه وثبات من طريقة online حيث يتم التعديل لكل مثال.
* أقل كلفة حسابيا لتعديل الأوزان من طريقة full-batch .حيث سنعالج كل أمثلة التدريب قبل أن نعدل الأوزان.
* حيث عمليات جداء الأشعة بالمصفوفات كما في المعادلة 34 و المعادلة 35 يمكن أن تصبح عبارة عن جداء مصفوفتين و التي يمكن أن تنفذ بشكل سريع على GPU.

*حيث يتم تعديل الأوزان في هذه الخوارزمية وفق العلاقة التالية:*

|  |  |  |
| --- | --- | --- |
| (39) |  |  |

حيث هو مشتق تابع الخطأ الذي نحاول تصغيره بالنسبة للأوزان , حيث تقوم هذه المعادلة تقوم بتعدل الأوزان في الاتجاه المعاكس للمشتق ، معدل التعلم و الذي يسمى في بعض المراجع حجم الخطوة ,هو **hyperparameter** يتحكم بحجم التحديث على الأوزان.

### 5.3.1مراجعة رياضية

لتبسيط العلاقات الرياضية في حساب المشتقات الجزء التالي يوضح بعض العلاقات و الملاحظات في حساب تفاضل الأشعة و حساب المشتقات.

ليكن u,v شعاعين و لتكن جداء الشعاعين على مستوى العناصر و ليكن مصفوفة قطرية قطرها هو الشعاع عندئذ يمكن أ، نكتب العلاقة التالية:

|  |  |  |
| --- | --- | --- |
| (40) |  |  |

***مبرهنة1:***  *ليكن لدينا الخطأ و التي لدينا مشتقه* بالنسبة للشعاع و ليكن ، فإن المشتقات للخطأ بالنسبة لشعاع و المصفوفة تعطى بالعلاقات:

|  |  |  |
| --- | --- | --- |
| (41) |  |  |
| (42) |  |  | |

البرهان ليكن

ليكن السطر ith من المصفوفة و ليكن  *العنصر ith من الشعاعين ، وليكن المشتقات لتابع الخطأ بالنسبة*  و بالتالي يكون لدينا

**نتيجة 1:***عندما تكون هو تابع المطابق [[4]](#footnote-6) يصبح يكون لدينا*

|  |  |  |
| --- | --- | --- |
| (43) |  |  |
| (44) |  |  | |

***ملاحظة 1:*** *ليكن أشعة بحيث ، وليكن مشتقات تابع الخطأ بالنسبة لهذه الأشعة*

|  |  |  |
| --- | --- | --- |
| (45) |  |  |
| (46) |  |  | |

**ملاحظة 2:** حالة خاصة من الملاحظة السابقة عندما يكون التابع المطابق  *يكون لدينا :*

|  |  |  |
| --- | --- | --- |
| (47) |  |  |
| (48) |  |  | |

### 5.3.2 الانتشار الخلفي لخطة زمنية واحدة

لحساب مشتقات تابع الخطأ ، يجب أن نكون قادرين على حساب المشتقات لخطوة الزمنية الواحد حيث هذه المشتقات تحسب بالنسبة أوزان الشبكة و بالإضافة نقوم بحساب المشتقات بالنسبة للدخل نرمز هذه المشتقات بـ و نقوم بتعريف المشتقات الوسيطة حيث مستعملة في المعادلتين 35 و 36 بدأ من الخطاء حيث نستعمل backpropagation في الهيكلية (Goller و Kuchler 1996)حيث نقوم بحساب المشتقات واحد كل مرة بالترتيب التالي

*من المعادلة* 37 *وباعتبار*  يكون لدينا:

|  |  |  |
| --- | --- | --- |
| (49) |  |  |

حساب المشتق من أجل كل بعد من أبعاد شعاع الخرج

|  |  |  |
| --- | --- | --- |
| (50) |  |  |

يمكن تمثيل المشتق السابق كشعاع يعطى بالعلاقة :

|  |  |  |
| --- | --- | --- |
| (51) |  |  |

حيث هو التوزيع الاحصائي المعرف بالمعادلة 2.5 التي في طور الانتشار الأمامي و بالتالي يمكن إعادة استعماله في طور الانتشار الخلفي. حيث هوشعاع one-hot مع القيمة 1 في الموضع و بتطبيق **نتيجة 1** حيث من المعادلة 35 يمكن استنتاج :

و بالتالي يكون مشتق تابع الخطأ بالنسبة للحالة الخفية و بالنسبة للأوزان الحالة الخفية و الخرج يعطى بالعلاقة :

|  |  |  |
| --- | --- | --- |
| (52) |  |  |
| (53) |  |  | |

وبالتالي قد استنتجنا جزء خوارزمية الانتشار الخلفي والتي يمكن تطبيقها على أي وحدة خفية في شبكات RNN أو وحدات LSTM التي سنقوم بتوصيفيها في الفصل التالي.

**الانتشار الخلفي في شبكات RNN العادية**

للتبسيط يمكن فرض المتحولات التالية و و بالتالي صيغة شبكات RNN في المعادلة رقم 34 تصبح

|  |  |  |
| --- | --- | --- |
| (54) |  |  |

و بتطبيق المبرهنة رقم 1 يصبح لدينا:

|  |  |  |  |
| --- | --- | --- | --- |
| (55) |  | |  |
| (56) |  |  | |

وهذه هي أحد الطرق المتبعة لتحسين الأداء على المسرع الرسومي GPU و ذلك بانشاء مصفوفات و أشعة كبيرة مثل من المعادلة 55 , و المعادلة 56 يمكن حساب المشتقات التالية

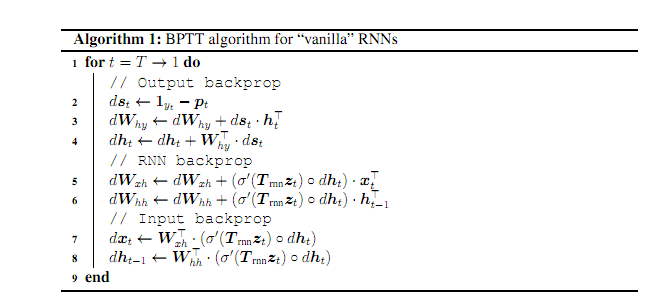
1. مشتق شعاع التضمين embedding و التي تستعملها لتحديث أوزان مصففة الـembedding .
2. المشتق بالنسبة للحالة الخفية السابقة، والتي نحتاجها لمعالجة الانتشار الخلفي في الزمن و التي سنتحدث عنها لاحقاً
3. الأوزان بين دخل شبكات RNN والطبقة الخفية.

### 5.3.3الانتشار الخلفي في الزمن Backpropagation Through Time (BPTT)

و بما أنه عرفنا خطوات خوارزمية الانتشار الخلفي من أجل خطوة زمنية واحدة، يمكننا الانتقال لخوارزمية الانتشار الخلفي (Rumelhart و McClelland، On learning the past tenses of English verbs. 1985) (Werbos و others 1990) و بالاعتماد على (Sutskever، Training recurrent neural networks 2013) يمكن تلخيص خوارزمية الموضحة بالخوارزمية 1 بالخطوات التالية

تلخيص خطوات BPTT بالخطوات التالية :

1. السطور 3,5,6,7 تجمع المشتقات وفق أوزان شركة خلال الزمن.
2. حيث في السطر رقم 7 تشير إلى مشتقات الكلمات المستعملة في هذه الـmini-batch و التي أبقيتها غير مجمعة و ذلك لتحديث الأوزان [[5]](#footnote-7)
3. السطر رقم4 يقوم بتجميع المشتقات من أجل الحالة الداخلية و هذا المشتقات يمكن اعتبار هذا المشتق ناتج من طرفين أولا الخطأ من اللحظة و الخطأ الناتج من اللحظة و التي تم حسابها في السطر رقم 8.



خوارزمية 1 الانتشار الخلفي خلال الزمن لشبكات RNN البسيطة

**الذاكرة طويلة قصيرة المدى Long Short-Term Memory (LSTM)**

وعلى الرغم من أن حساب المشتقات واضح نسبياً إلا أن تدريب شبكات RNN صعب جداً بسبب الطبيعة التكرارية الغير خطية لشبكات RNN، أحد المشاكل الرئيسية الشائعة في شبكات RNN تظهر عند التعامل مع السلاسل الطويلة نسبياً هي مشكلة الاضمحلال Vanishing، الانفجار exploding (Bengio, Simard, et al. 1994) وبتعريف مختصر:

الانفجار exploding وتحدث عندما تصبح الأوزان كبيرا جدا وتتزايد بشكل كبير جداً في الانتشار الخلفي مع الزمن مما يجعل التعليم غير مستقر.

أما الاضمحلال vanishing فتحدث عندما تنحدر الأوزان بسرعة كبيرة نحو الصفر بحيث تصبح شبكات RNN عاجزة عن كشف العلاقات في السلاسل الطويلة.

**نوقشت هذه المشكلة من قبل العديد من الباحثين** (Bengio، Simard، وآخرون 1994)(Hochreiter و Schmidhuber 1997)(Martens و Sutskever 2011)(Pascanu, Mikolov and Bengio 2013) **المشكل الحقيقية لهاتين المشكلتين تمكن في السطر 8 من الخوارزمية 1 و التي يمكن إعادة صياغتها** بالشكل التالي *انظر الملاحظة* 1*، و لفهم سلوك شبكات RNN وذلك بفرض أن السطر* 4 *غير موجود أي نفترض الوحدات الخفية لا تساهم بالخطأ .*

*فأن الإشارة المنتشرة خلفياً من الحالة الخفية الحالية بعد خطوة زمنية* ستصبح , بفرض أن التابع غير خطي كتابع sigmoid او tanh سيصبح علينا التعامل مع جداء المصفوفة التكرارية وفق الخطوات الزمنية. و هذا يؤدي ان سلوك شبكات RNN محكوم محدد بخصائص المصفوفة التكرارية , و بمساعدة بعض حيل الجبر الخطي تبين أن إذا كانت أكبر قيمة لـeigenvalue للمصفوفة كبيرة كفاية فإن الانفجار سيحدث على الأغلب، و على النفيض من هذا إذا كانت أكبر قيمة لـeigenvalue للمصفوفة أصغر من حدا معين فإن الاضمحلال سوف يحدث كما تم شرحه بالتفصيل من قبل (Pascanu، Mikolov و Bengio 2013).

**تقليم المشتقات :Gradient Clipping** في الواقع يمكن التعامل مع مشكلة انفجار المشتقات وذلك بتطبيق عدة طرق مختلفة من طرق تقليم المشتقات، الطريقة الأولى تم اقتراحها من قبل (T. Mikolov 2012)و ذلك بتطبيق معامل تقليم على مستوى العنصر ،حيث خلال كل خطوة زمنية في مرحلة الانتشار الخلفي ، أية عنصر من أكبر من حد موجب معين سيصبح أو أصغر من سوف يصبح *.*

*يمكن إجراء تقليم على مستوى المشتق هذه الطريقة مقترحة من قبل* (Razvan, Tomáš and Yoshua 2013) *بالحصول على شعاع المشتق الأخير المحسوب المحسوب عند كل mini-batch فإذا كان طويلته أكبر من حد معين* ،فنقوم بإعادة تقيس هذا شعاع المشتقات وفق العلاقة و هذه الطريقة هي الأكثر استعمالاً كما يمكن استعمالها مع الطريقة الأولى.

**الذاكرة طويلة قصيرة المدى :Long Short-Term Memory (LSTM)**لكن مشكلة الاضمحلال من ناحية أخرى أكثر صعوبة للتعامل معها هنالك عدة طرق للتعامل معها تجاوز الوصلات (Alexander Waibel و J.Lang 1990) (Lin، وآخرون 1996)، البنية الهرمية (El Hihi و Bengio 1996)، Leaky Integrator (Jaeger, et al. 2007) ، و العديد الطرق الأخرى التي وضحها (Bengio, Boulanger-Lewandowski and Pascanu, Advances in optimizing recurrent networks 2013).

من بين هذه الطرق Long short-term Memory (LSTM) تم اقتراحها من قبل (Hochreiter و Schmidhuber 1997) (Gers, Schmidhuber and Cummins 2000) و يبدو أنه من أنها افضل الطرق لحل مشكلة الاضمحلال .

الفكرة الرئيسية من LSTM هو تطعيم شبكات RNN بذواكر خطية تسمع للمشتق بالانسياب بشكل ناعم خلال الزمن، كما هنالك وحدات بوابات و التي تتحكم في مدى استخدام شبكات RNN للذاكرة ،(بوابات للنسيان) حيث تستقبل إشارات الدخل (بوابات ا لدخول) و تستخلص المعلومات (بوابات الخرج) في كل خطو زمنية

هنالك العديد من طرق تنفيذ خلايا LSTM تختلف فيما إذا كان الانحياز مستعمل أو أية انحياز مستعمل كيفية بناء البوابات وبيّن (Greff, et al. 2017) (Jozefowicz, et al. 2016) أن هذا الاختلاف غير مهمو لذلك سنلتزم بتعريف واحد لهذه البوابات (Zaremba، Sutskever و Vinyals 2014)

قبل البحث في التفاصيل سنقوم بشرح عن بنية LSTM بشكل تدريجي، يمكن بناء ذاكرة بسيطة كالتالي:

|  |  |  |
| --- | --- | --- |
| (57) |  |  |
| (58) |  |  |

هذه البنية يمكن اعتبارها أحد أشكال leaky integration التي وصفها (Sutskever، Training recurrent neural networks 2013) (Bengio، Boulanger-Lewandowski و Pascanu، Advances in optimizing recurrent networks 2013) حيث إنها مكافئة , تدريب هذه الشبكة للسلاسل الطويلة سهل نسبياً حيث من أجل كل المسارات الأسية هنالك طريق وحيد يعبر وحدات الذاكرة و هذا يضمن أن لا تضمر بسبب في هذا المسار .

هذه البنية غير مناسبة لأنه لدخل معين (الكلمات الوظيفية – أو علامات الترقيم )لا علاقة لها بالمهمة الحالية و يجب أن يخفض وزنها، أحيانا نريد أن نعيد تصفير الذاكرة مثل بداية الجملة في مقطع نصي، ولزيادة المرونة و فعالية هذه البنية نضيف بوابات للنسيان و الدخل و الخرج كالتالي:

|  |  |  |
| --- | --- | --- |
| (59) |  |  |
| (60) |  |  | |

يلاحظ أنه في المعادلة 60 أن خلية الذاكرة قد مُررت لتابع غير خطي قبل استعمال بوابة الخرج لاستخلاص LSTM. و للتأكد من أن البوابات قابلة لتكيف نبنيهم بناء على المعلومات المعطات من الدخل الحالي ،و الحالة الخفية السابقة و نريد البوبات أن تكون في المجال و بالتالي تابع السيغمويد سوف يتم استعماله في البوابات،و بناء على ما تحدثنا عنه سابقا يمكن توصيف LSTM كما هو موضح في (Zaremba، Sutskever و Vinyals 2014)حيث التابع هو التالبع tanh:

|  |  |  |
| --- | --- | --- |
| (61) |  |  |
| (62) |  |  |
| (63) |  | |  |

و كما في المعادلة 54 يمكن أن يكون الحساب أكثر فعالية على المحول الرسومي GPU ففي المعالدلة 61 جمعنا 8 مصفوفات إلى مصفوفة واحدة كبيرة و التي دعوناها و باعتبار ،نقوم أولاً بالجداء و بعدها نطبق توابع خطية مختلفة لاجزاء الخرج الموافقة، و لحساب الانتشار الخلفي بشكل أسهل نعيد كتابة المعادلة 61 لتصبح:

|  |  |  |
| --- | --- | --- |
| (64) |  |  |
| (65) |  |  | |
|  |  |  | |

حيث g تابع غير خطي يطبق على كل عنصر على حدا و لم نعرف g بشكل صريح لأنه سيكون tanh للقسم الشعاعي الموافق لـ و sigoid للبقية.

**LSTM Training** في مسار تدريب شبكات LSTM هنالك العديد من العناصر التي هي نفسها أو مشابهة لتدريب شبكات RNN.

أولا سنذكر الخلافات الرئيسية أولا في التابع التكراري (recurrence function) لا يوجد فقط الحالة الخفية و لكن هنالك خلايا الذاكرة كدخل و خرج :

|  |  |  |
| --- | --- | --- |
| (66) |  |  |

و في هذه الحالة التابع *f*  معرف بالمعادلات 61,63 و حالما يتم حساب فإن آلية التوقع نفسها كما في شبكات RNN المعرفة بالمعادلة 35,37 هدف التدريب في المعادلة 38يبقى بدون تغيير أيضأ.

**LSTM Backpropagation** بما ألية التوقع نفسها فان الانتشار الخلفي لـLSTM هي نفسها في شبكات RNN حتى المعادلتين 52،53و التي تحسب و على التوالي

و بما أنه أصبح لدينا يمكننا الآن التراجع للحصول على المشتقات الأخرى بدأ من المعادلة 63 و بتطبيق **الملاحظة 2** يصبح لدينا :

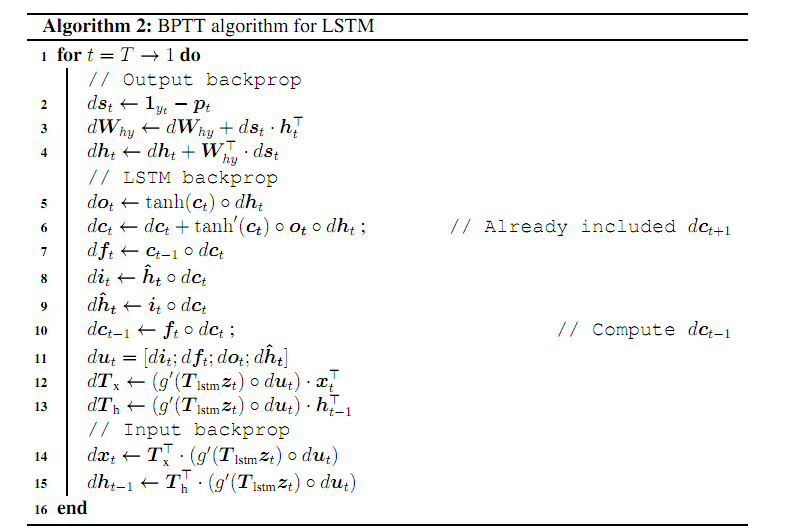
|  |  |  |
| --- | --- | --- |
| (67) |  |  |
| (68) |  |  |
| (69) |  | |  |
| (70) |  | |  | |

و قبل الانتشار الخلفي في المعادلة (62 يجب ان تذكر تحديث المشتق بالمشتق القادم من و التي تم اجرائها بالسطر 6 و 10 من الخوارزمية ،ليدنا المشتق و بتطبيق الملاحظة رقم 3 يمكن استنتاج

|  |  |  |
| --- | --- | --- |
| (71) |  |  |
| (72) |  |  |
| (73) |  | |  |
| (74) |  | |  | |

وبعد حساب كل المشتقات يمكن جمعها لكتابة خوارزمية الانتشار الخلفي لشبكات LSTM

خوارزمية 2 خوارزمية الانتشار الخلفي في الزمن لشبكات RNN بخلايا نوع LSTM



# .6الترجمة الآلية باستخدام الشبكات العصبية Neural Machine Translation (NMT)

كنا قد تكلمنا عن النماذج اللغوية باستخدام شبكات RNN يمكن اعتبار الترجمة الآلية باستخدام الشبكات العصبية نموذج لغوي شرطي بالنسبة للجملة المصدرية،NMT تسعى إلى نمذجة الاحتمال الشرطي وذلك بترجمة الجملة المصدر للجملة الهدف ، و هي تقوم بتحقيق هذا الهدف من خلال هيكلية encoder-decoder (Sutskever، Vinyals و Le، Sequence to sequence learning with neural networks 2014) (Kalchbrenner و Blunsom، Recurrent continuous translation models 2013، Chorowski، Bahdanau، وآخرون، End-to-end continuous speech recognition using attention-based recurrent nn: First results 2014) حيث يعمل encoder المرمز على حساب تمثيل لكل جملة مصدرية، وبناء على هذا التمثيل لجملة المصدر يقوم decoder فاك الترميز بتوليد ترجمة ، حيث يقوم بتوليد كلمة هدف في كل مرة ،حيث يمكن أن نفكك لوغاريتم الاحتمال الشرطي السابق إلى المعادلة التالية:

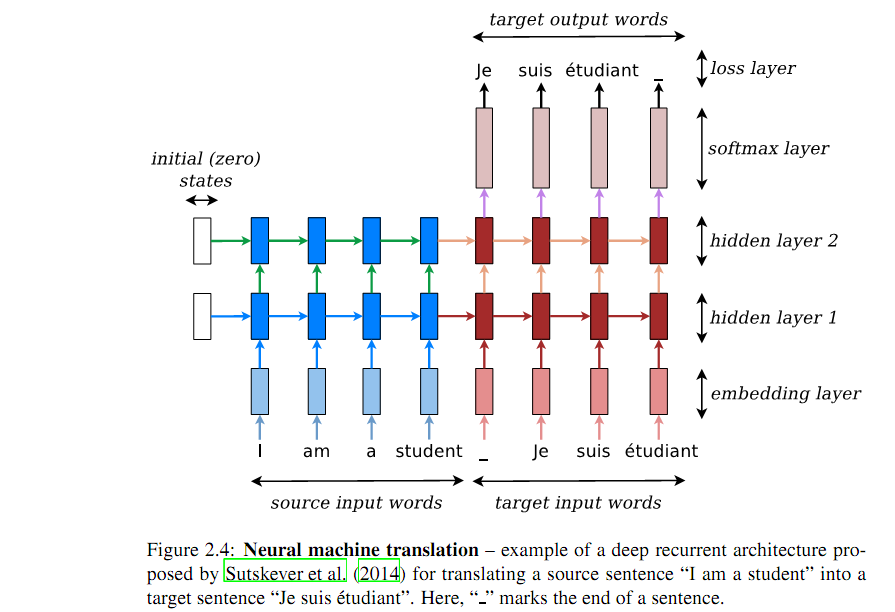
|  |  |  |
| --- | --- | --- |
| (75) |  |  |

نماذج NMT تتابين في بنيتها، نظرا للطبيعة المتسلسلة للبيانات المستخدمة فتعد شبكات RNN أكثر استخداما و تقريبا معظمها تستخدم بنية encoder-decoder نماذج RNN المستعملة تختلف بالأمور التالية

1. الاتجاه: وحيدة الاتجاه ثنائية الاتجاه
2. العمق: شبكة من طبقة و احدة أو عميقة من عدة طبقات
3. النوع: هل هي شبكة RNN عادية أو (Hochreiter و Schmidhuber 1997) LSTM أو gate recurrent unit (GRU) (Chorowski، Bahdanau، وآخرون، End-to-end continuous speech recognition using attention-based recurrent nn: First results 2014).

بالنسبة للمرمز encoder أية بنية يمكن استخدامها حيث (Kalchbrenner and Blunsom, Recurrent continuous translation models 2013) استخدم convolutional neural network لترميز المصدر. بينما الخيارات لفك الترميز تكون محدودة أكثر حيث المطلوب هو توليد جملة الترجمة، أكثر الطرق المستعملة هي RNN وحيدة الاتجاه والتي تبسط خوارزمية فك الترميز beam-decoding.

سنستحدث في الجزء التالي عن البنية التي اقترحها (Sutskever, Vinyals and Le, Sequence to sequence learning with neural networks 2014) و التي هي عبارة عن شبكة عميقة متعددة الطبقات وحيدة الاتجاه وتستخدم وحدات LSTM للوحدات الخفية ، من وجهة نظر عليا هذا النموذج يتألف من شبكتي RNN، في **مرحلة التدريب** تقوم شبكة المرمز بمعالجة الجملة المصدرية بدون أن تقوم بأية توقع , اما شبكة فك الترميز فإنها تعالج الجملة الهدف بالإضافة إلى توقع الكلمة الهدف التالية.



و بالتفصيل أكثر يتلقى المرمز الجملة المصدر ثم علامة نهاية الجملة "-" و التي يعني الانتقال من عملية الترميز إلى فك الترميز و بعدها يتم تقديم الجملة الهدف إلى فاك الترميز حيث تعطى الجملة الهدف للدخل و خرج الشبكة لحساب خطأ الشبكة حيث أن الجملة الهدف عند الخرج مزاحة للخلف بمقدار واحد، هذه الكلمات المتقطعة في الجملة المصدرية أو الهدف حيث يقوم النموذج بالبحث عن ترميز الكلمات في الطبقةembeddings المصدر و الهدف للحصول على التمثيل المناسب للكلمة، أوزان الطبقة يمكن تهئيتها خلال التدريب، كما يمكن تدريبها بشكل منفصل كـword2vec (Mikolov et al.2013) Glove (Pennington et al., 2014) .

بعد الحصول على تمثيل الكلمات فإن هذا التمثيل يتم تغذيته للشبكة الرئيسية, و التي تتألف من شبكتي RNN من عدة طبقات مربوطتين مع بعضها، يمكن لشبكتين ان تتشاركا نفس الأوزان لكن التجربة بينت أن الحصول على أوزان مختلفة مناسب أكثر وتقلل من overfitting .

يستعمل المرمز شعاع صفري كحالته البدائية، بينما فالك الترميز من ناحية أخرى بالحاجة إلى الوصول إلى بعض معلومات المصدر أحد الطرق البسيطة هو بتهيئة الحالة البدائية لفاك الترميز بالحالة النهائية للمرمز كما هو موضح بالشكل 2.4 حيث الحالة الخفية للمرمز عند الكلمة "student" تمرر لفاك الترميز.

الأوزان الأمامية (العامودية) تربط الوحدات الخفية من الطبق السفلى إلى الطبقة العليا في حين أن الأوزن الدورانية تنقل (الأفقية ) تنقل المعرفة من الخطوة السابقة للخطوة التالية ,في هذه البنية يوجد 4 مجموعات أوزان مختلفة لـLSTM موضحة بالمعادلة 64 وهي أوزان طبقتين في الـencoder و أوزان طبقتين في الـdecoder , و في الخطوة الأخيرة فأن الحالة الخفية في الطبقة العليا تحول إلى توزيع إحصائي على مجموعة الكلمات وفق تابع softmax وفق المعادلتين 36,35.

## 6.1التدريب:

تدريب شبكات الترجمة الآلية مشابه لتدريب النماذج اللغوية لشبكات RNN و التي تم مناقشتها في الفقرات السابقة غير أنه يجب التعامل مع الشرط بالنسبة لجملة المصدر و بالتالي يصبح تابع الخطأ

|  |  |  |
| --- | --- | --- |
| (76) |  |  |

*حيث يشير إلى المدونة المتناظرة و المستخدمة للتدريب للجمل المصدرية و الهدف . بالنسبة المقترحة لبنية الشبكة العصبية المقترحة للترجمة فإن حساب الخطأ لـلزوج الجملتين في طور الانتشار الأمامي تقريباُ هو نفس حساب الخطأ في شبكة* RNN *بالجملة فقط .الاختلاف الوحيد هو أنه نحن بحاجة إلا إنشاء تمثيل للجملة المصدر أولاً و ذلك لتهيئة فاك الترميز عوضاً عن البدء بالحالة الصفرية.*

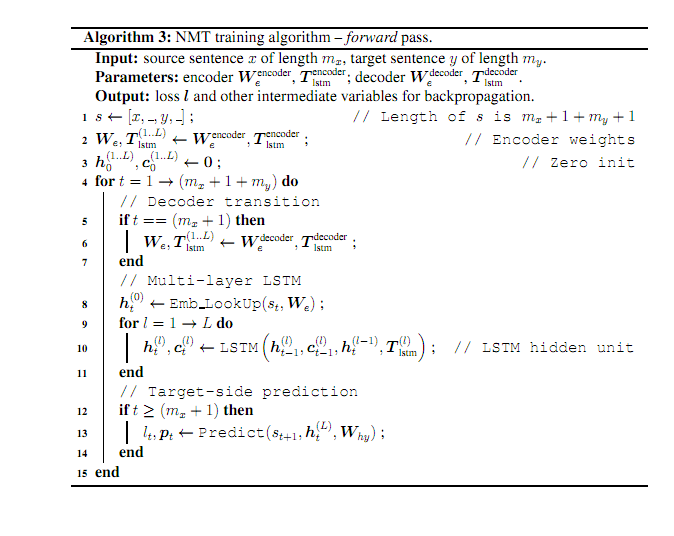
*في طور مرحلة الانتشار الخلفي، حساب المشتقات لفاك الترميز هو نفسه ما تم شرحه في* الخوارزمية 2*حيث مشتق الحالة الخفية في فاك الترميز يمرر خلفيا للمرمز وبعدها نستمر في الانتشار الخلفي عبر المرمز وبنفس الأسلوب المتبع في فالك الترميز ولكن بدون حساب خطأ التوقع.*

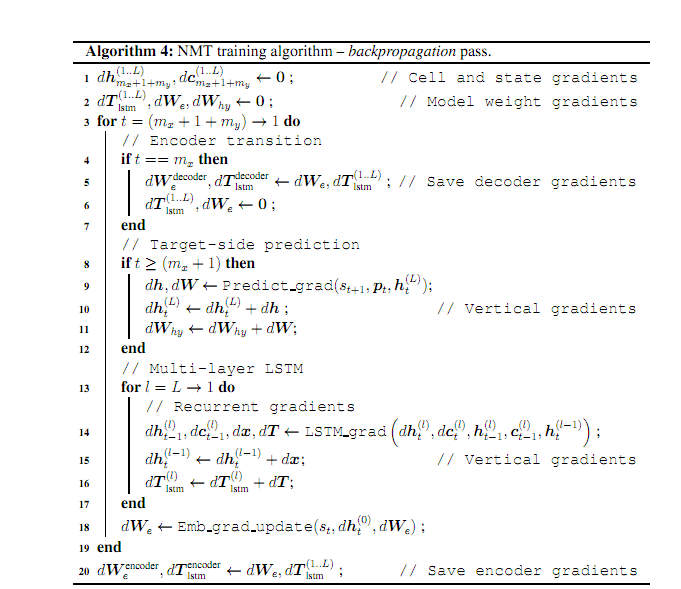
## 6.1.1 طور الانتشار الأمامي

في الخوارزمية رقم 3 توضح طور الانتشار الأمامي في نموذج NMT مع بنية متعددة الطبقات من وحدات LSTM.و بما أن المرمز و فاك الترميز يشتركان بالعديد من العمليات جمعنا جملة المصدر و طولها و جملة الهدف و طولها مع رمز نهاية الجملة "-" مع بعضهم لتشكيل سلسلة الدخل كما هو مبين في السطر رقم 1.نبدأ أولاً بأوزان المرمز و بالحالة البدائية مسندة إلى الحالة الصفرية (السطرين 2-3) و في الخطوة (السطر 5) تنتقل الخوارزمية إلى وضع فك الترميز. الخوارزمية نفسها المستعمل *لوحدات* LSTM *في الترميز و فك الترميز (الأسطر 8-11) حيث يتم البحث عن الـembeding من أجل الدخل ؛ و بعد هذا يتم حساب الحالة الخفية وقيم خلايا الذاكرة من الطبقة السفلى إلى الطبقة العليا (الطبقة الطبقة L).*

*في السطر 10 يتم تخزين قيمة LSTM* *حسب الصيغة المبينة في المعادلتين* 61*,*63 *والتي يمكن بسهولة استبدالها بوحدات خفية أخرى كوحدات RNN أو GRU , شيء أخير في طرف فاك الترميز الحالة الخفية العليا تستخدم لتوقع الرمز التالي*  (السطر 13) و بعدها حساب الخطأ و التوزريع الأحصائي يتم حسابه بناء على المعادلة 35,36.

خوارزمية 3 تدريب شبكة الترجمة الآلية (طور الانتشار الأمامي)





## 6.1.2 طور الانتشار الخلفي

و هي موضحة بالخوارزمية رقم 4، و بالتدقيق أكثر في الخوارزمية نلاحظ تشابه كبير مقارنة بمرحلة الانتشار الأمامي إلا أن العملية أصبحت معكوسة.

أولاً نهيئ المشتقات في الطبقات التكرارية عند الخطوة الزمنة الأخيرة (السطر 1) بالإضافة إلى مشتقات الأوزان عند فالك التشفير (السطر 2) بالقيمة 0. في الخطوة الزمنية نتحول إلى مرحلة فك الترميز و ذلك بحفظ مشتقات الـLSTM و الـembedding الخاصة بفاك الترميز (السطر 5) و نبدأ بعدها بتجميع المشتقات لأوزان المرمز (السطر 6).

و بفضل خوارزمية الانتشار الخلفي لشبكات LSTM يمكن تبسيط حساب المشتقات لـNMT (الأسطر 8-18) و ذلك بالإشارة إلى عنصرين

1. Predic\_grad (الأسطر 2-4 من الخوارزمية 2)و التي تحسب مشتقات تابع الخطأ بالنسبة للحالة الخفية في الطبقة العليا و أوزان تابع softwax
2. LSTM\_grad (الأسطر 5-15 من الخوارزمية 2) و التي تحسب المشتقات بالنسبة لدخل LSTM و أوزان الـLSTM في كل طبقة

(السطرين 10،15) من الخوارزمية 4 نضيف المشتقات المتجهة عمودياً أما من تابع الخطأ أو من طبقات LSTM العليا للمشتقات التي تحتها.

# .7آلية الانتباه (Attention Mechanism)

على الرغم من أن الترجمة الآلية بالاعتماد على الشبكات العصبية NMT قد حصلت على أفضل النتائج في مجال الترجمة الآلية , فإنه من الصعب على شبكات NMT التعامل مع الجمل الطويلة (Bahdanau, Cho and Bengio 2015) . أحد الطرق الفعالة لمعالجة هذه المشكلة من خلال آليى الانتباه , ة التي لاقت اهتمام كبير مؤخرا في تدريب الشبكات العصبية لكي تتعلم الملائمة بين كيفيات مختلفة مثل (الصور , أفعال نظام تحكم (Mnih, et al. 2014) بين المقاطع الصوتية النص في أنظمة التعرف على الكلام (Chorowski, Bahdanau, et al., End-to-end continuous speech recognition using attention-based recurrent nn: First results 2014) بين الخصائص البصرية لصورة ووصفها النصي في أنظمة توليد تعريف الصور (Xu, et al. 2015) ،

وفي مجال (Bahdanau, Cho and Bengio 2015) NMT قد استعمل آلية الانتباه للترجمة و ترتيب الكلمات ,هنالك نوعين رئيسين من النماذج المعتمدة على الملاحظة . النموذج العام حيث كل كلمات الجملة المصدر يتم ملاحظتها ،النموذج المحلي حيث مجموعة جزئية من كلمات الجملة المصدر يتم ملاحظتها في كل مرة ، و هذ النموذج الأخير يشايه ما استعمل (Bahdanau, Cho and Bengio 2015) و لكن ببنية أبسط .

## 7.1الأنظمة المبنية على الانتباه

على خلاف أنظمة NMT التقليدية (Kalchbrenner and Blunsom 2013) (Sutskever, Vinyals and Le 2014) (Chorowski, Bahdanau, et al. 2014) (Minh-Thang, et al. 2015) حيث يتم انشاء تمثيل وحيد لجملة الدخل يتم به تهيئة الحالة الداخلية لفاك الترميز, الفكرة من ألية الانتباه موضحة في (Bahdanau, Cho and Bengio 2015) حيث تمكننا هذه الآلية تنشأ "ذاكرة وصل عشوائي" للحالات الداخلية للمصدر التي تشير إلى تقدم عملية الترجمة .

الأنظمة لمختلفة التي تعتمد على آلية الانتباه التي يتم شرحها في هذا الفصل يمكن تصنيفها على صنفين الانتباه العام والانتباه المحلي الفرق بين هذه الصنفين هو أن الانتباه يكون أما مركزاً على كل مواضع الجملة المصدر أو أن يكون مركزاً على بعض المواقع في الجملة المصدر هذا الصنفين موضحين بالشكل 4.2 و 4.3.

الشيء المشترك بين هذين النوعين، عند كل خطوة زمنية في مرحلة فالك الترميز ،كلا النوعين يأخذان الحالة الخفية و الطبقة العليا من شبكة LSTM،لكن الهدف من آلية الانتباه هو إيجاد شعاع السياق و التي يمثل المعلومات المهمة في المصدر للمساعدة في توقع الكلمة الهدف الحالية . الفرق بين هذا النوعين هو كيفية إيجاد شعاع السياق و يتشاركان في الخطوات التالية، و بالتحديد بالنسبة للحالة الخفية و شعاع السياق في المصدر , نقوم بوصل الحالتين مع بعضهما و نقوم بحساب الحالة الداخلية لآلية الانتباه أو ما يسمى شعاع الملاحظات من العلاقة :

|  |  |  |
| --- | --- | --- |
| (77) |  |  |

وبعدها يمرر شعاع الملاحظات تمريره عبر طبقة softmax لإيجاد التوزيع الإحصائي المتوقع وفق العلاقة:

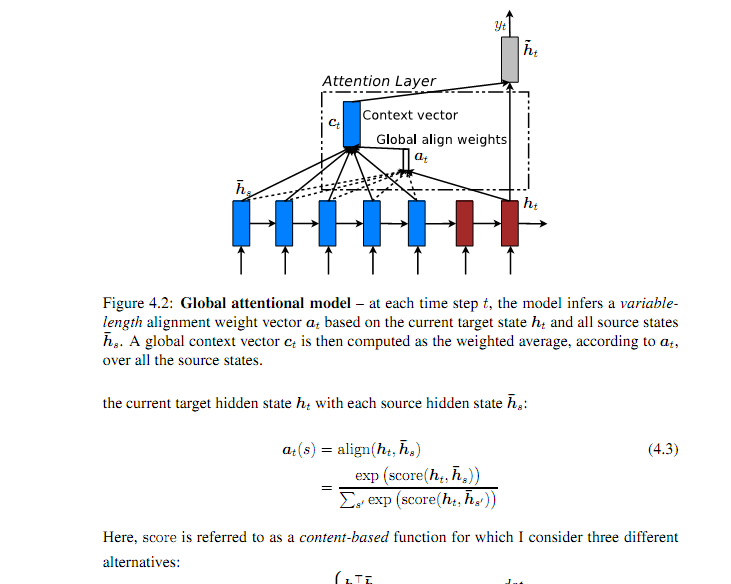
|  |  |  |
| --- | --- | --- |
| (78) |  |  |

## 7.2 حساب شعاع السياق

### الانتباه العام

في أنظمة الانتباه العام يتم اعتبار جميع الحالات الخفية للمرمز في الحسبان عند حساب شعاع السياق . في هذا النوع من النماذج يتم حساب شعاع ترتيب aligment vector ذو طول متغير و الذي حجمه يساوي عدد الخطوات الزمنية في المصدر، حيت يتم حساب هذا الشعاع بمقارنة الحالة الداخلية الحالية مع كل حالة داخلية في المرمز

|  |  |  |
| --- | --- | --- |
| (79) |  |  |



حيث score هو تابع معتمد على المحتوى و التي يمكن تعريفه بإحدى الطرق التالية

|  |  |  |
| --- | --- | --- |
| (80) |  |  |

في محاولتنا لبناء نموذج قائم على الانتباه نستعمل تابع تعتمد على الموضع حيث علامات الترتيب aligment score تعتمد على الحالة الخفية الهدف من العلاقة :

|  |  |  |
| --- | --- | --- |
| (81) |  |  |

و باعتبار أن شعاع الترتيب alignment vector هو وزن فإن شعاع السياق  *يمكن حسابه متوسط حسابي مثقل لكل الحالات الخفية المصدر [[6]](#footnote-8)*

*الطريقة المستعملة في هذا البحث ستكون مشابهة للمقاربة المستخدمة في* (Bahdanau, Cho and Bengio 2015) *إلا أنه هنالك عدة اختلافات رئيسية لتبسيط الطريقة الأصلية.*

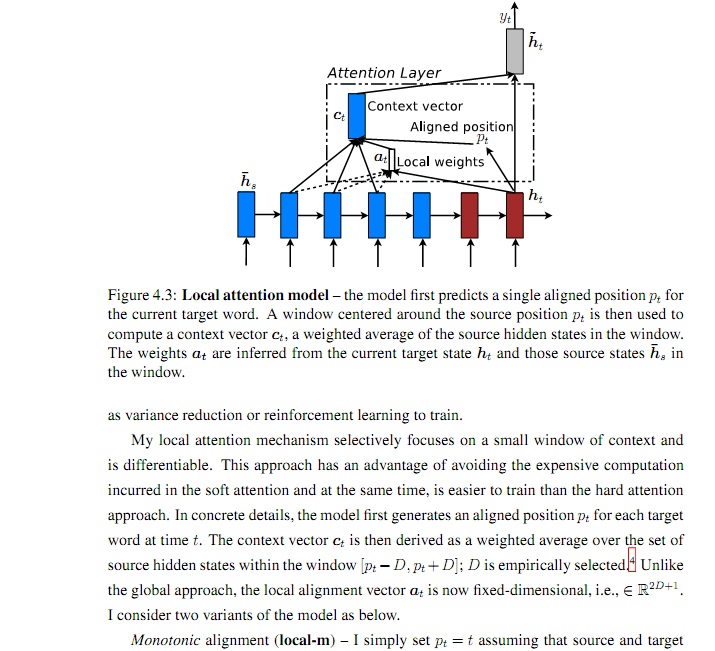
*أولا تم استعمال الحالة الخفية التي تمثل الطبقة العليل من LSTM في كلا من المرمز و فاك الترميز .كما هو موضح بالشكل 4.2* (Bahdanau, Cho and Bengio 2015) *استعمل وصل الحالتين الخفيتين الأمامية و الخلفية للشبكة RRN ثنائية الاتجاه و الحالة الخفية في الخرج حيث شبكة فاك الترميز تتألف من شكة RNN وحيدة الاتجاه .*

*ثانياً مسار الحساب المتبع حيث التوقع الناتج عن النموذج موضح بالمعادلة* 77*و المعادلة* 78 *و الشكل 4.2 و من ناحية أخرى* (Bahdanau, Cho and Bengio 2015) *انطلق من الحالة الخفية السابقة و التي بدورها خلال شبكة عميقة و طبقة لاختيار القيمة العظمى قبل ان تقم بالتوقع[[7]](#footnote-9)*

*أخيرا* (Bahdanau, Cho and Bengio 2015)*أجرى تجارب على تابع ترتيب واحد بينما يمكن ان تكون الخيارات الأخرى أفضل.*

### الانتباه المحلي

أحد سلبيات الانتباه العام بأنه يلاحظ كل كلمات المصدر لأجل كل كلمة في الهدف و التي هي عملية مكلفة و التي يمكن أن تكون غير مفيدة عند ترجمة الجمل الطويلة و لحل هذ المشكلة فإن آلية االانتباه المحلي تركز على مجموعة جزئية من كلمات الجملة المصدر عند توليد كل كلمة في الخرج.

هذا النموذج قد تم اقتراحه كحل وسط بين النماذج المرنة و الشديد للملاحظات و الذي اقترحه (Xu, et al. 2015)ليحل مشكلة توليد توصيف نصي للصور، حيث في بحثهم الانتباه المرن يشير إلى الانتباه العام حيث الأوزان يتم اسنادها بطريقة مرنة لكل مجموعات صور التدريب المصدر , الانتباه المرن يختار صورة في كل مرة ليركز الانتباه عليها , بينما هذه الطريقة أقل كلفة عند التشغيل و لكن الانتباه الشديد غير قابل للاشتقاق و يتطلب تقنيات معقدة كتقليل التباين variance reduction أو التعليم المعزز enforcment learning.

الانتباه المحلي الذي نستعمله يركز على نافذة صغيرة من السياق و قابل للاشتقاق ,هذه المقاربة تتجنب الحسابات المكلفة التي يسببه الانتباه المرن و هي أسهل من التدريب الانتباه الشديد

و بالتحديد يقوم النموذج بتوليد موضع مرتب لكل كلمة في الخرج في الخطوة و بعدها شعاع السياق يتم حسابه كمتوسط موزون للحالات الخفية ضمن نافذة حيث يتم اختيارها تجريببيا[[8]](#footnote-10)

و على خلاف الانتباه العام شعاع الترتيب المحلي  *له أبعاد ثابتة يوجد نوعين من الترتيب*

*Monotonic aligment local-m: حيث و تفترض أن الجملة المصدر و الهدف لهما نفس الترتيب حيث شعاع الترتيب aligment vector معرف بالمعادلة 4.3[[9]](#footnote-11)*

*Predictive aligment local-p يتم حساب فيه المواقع المتقابلة كالتالي :*

|  |  |  |
| --- | --- | --- |
| (82) |  |  |

*حيث و هي برمترات (أوزان) النموذج التي يجب تعلمها خلال التدريب لتتوقع المواضع ،S طول الجملة الأصلية و بالتالي فأن ، و لكي تناسب مواضع الترتيب قرب اعتبرنا وجود توزيع احصائي غاوصي مركزه و بالتالي فإن أوزان الترتيب يمكن تعريفها بالعلاقة:*

|  |  |  |
| --- | --- | --- |
| (83) |  |  |

*سنستعمل نفس تابع الترتيب في المعادلة* 79*و الانحراف المعياري تم وضعه تجريبيا ليسلو*  مع الملاحظة أن عدد حقيقي بينما هو رقم صحيح مع نافذة مركزها [[10]](#footnote-12)

1. الترجمة ال
2. مبادئ الترجمة الإحصائيةStatistical MT Preliminaries
3. النماذج اللغوية(تعريفها و أهميتها)
4. أنواعها النماذج اللغوية
5. نماذج N-GRAM
6. النماذج المعتمدة على القواعد اللغوية:
7. النماذج اللغوية الخطية اللوغاريتمية
8. النماذج اللغوية و تعليم الآلة

# الشبكات العصبية

1. تعريف
2. تدريب الشبكات العصبية (back propagation)
3. شبكات Recurrent Neural Networks
4. تدريب شبكات RNN باستخدام الانتشار العكسي
5. مراجعة رياضية
6. عيوب شبكات RNN
7. خلايا LSTM

# الترجمة الآلية باستخدام الشبكات العصبية Neural Machine Translation (NMT)

1. نموذج encoder-decoder
2. طور الانتشار الأمامي
3. طور الانتشار الخلفي
4. تدريب الشبكة
5. آلية الانتباه و انواعه

# تقيم الترجمة الآلية

1. معايير تقيم الترجمة الآلية
2. معيار BLEU

# القسم العملي

## البيانات المستخدمة لتدريب الشبكة العصبية

## الترجمة من العربي إلى الانكليزي

1. بنية الشبكة العصبية التي تم استخدامها و بارومتراتها
2. تدريب الشبكة
3. النتائج التي تم الحصول عليها

## الترجمة من العربي إلى الانكليزي

1. بنية الشبكة العصبية التي تم استخدامها و بارومتراتها
2. تدريب الشبكة
3. النتائج التي تم الحصول عليها

## مقارنة مع دراسات أخرى

## خاتمة

# مراجع

Adam L. Berger، Vincent J. Della Pietra، و Stephen A. Della Pietra. "A maximum entropy approach to natural language processing." *Computational linguistics* (MIT Press) 22 (1996): 39-71.

Adam Pauls، و Dan Klein. "Faster and smaller n-gram language models." *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1.* 2011. 258-267.

Brown, Peter F., Peter V. Desouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. "Class-based n-gram models of natural language." *Computational linguistics* (MIT Press) 18 (1992): 467-479.

Christoph Goller، و Andreas Kuchler. "Learning task-dependent distributed representations by backpropagation through structure." *Proceedings of International Conference on Neural Networks (ICNN'96).* 1996. 347-352.

Daniel Marcu، و Daniel Wong. "A phrase-based, joint probability model for statistical machine translation." *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002).* 2002.

David E. Rumelhart، Geoffrey E. Hinton، و Ronald J. Williams. "Explorations in the microstructure of cognition." *Parallel Distributed Processing*، 1986.

David E. Rumelhart، و James L. McClelland. "On learning the past tenses of English verbs." Tech. rep.، CALIFORNIA UNIV SAN DIEGO LA JOLLA INST FOR COGNITIVE SCIENCE، 1985.

David M. Blei، Andrew Y. Ng، و Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3 (2003): 993-1022.

David Talbot، و Thorsten Brants. "Randomized language models via perfect hash functions." *Proceedings of ACL-08: HLT*، 2008: 505-513.

Diederik P. Kingma، و Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980*، 2014.

Dzmitry Bahdanau، Kyunghyun Cho، و Yoshua Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate." *ICLR*، 2015.

Felix A. Gers، Jürgen Schmidhuber، و Fred Cummins. "Learning to Forget: Continual Prediction with LSTM." *Neural Computation* (MITP) 12 (2000): 2451-2471.

Franz Josef Och، و Hermann Ney. "Discriminative training and maximum entropy models for statistical machine translation." *Proceedings of the 40th annual meeting on association for computational linguistics.* 2002. 295-302.

Geofrey Hinton Kiyohiro Shikano Alexander Waibel، و Kevin J.Lang. *Readings in speech recognition. chapter Phoneme Recognition Using Time delay Neural Networks.* 1990.

Graham Neubig and Chris Dyer. "Generalizing and hybridizing count-based and neural language models." *the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP).* 2016.

Herbert Jaeger، Mantas Lukoševičius، Dan Popovici، و Udo Siewert. "Optimization and applications of echo state networks with leaky-integrator neurons." *Neural networks* (Elsevier) 20 (2007): 335-352.

I.H. Witten ، و T.C. Bell. "The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression." *Information Theory, IEEE Transactions*، 1991: 1085-1094.

Ilya Sutskever. *Training recurrent neural networks.* University of Toronto Toronto, Ontario, Canada، 2013.

Ilya Sutskever، Oriol Vinyals، و Quoc V Le. "Sequence to sequence learning with neural networks." *Advances in neural information processing systems.* 2014. 3104-3112.

Ilya Sutskever، Oriol Vinyals، و Quoc V. Le. "Sequence to sequence learning with neural networks." *Advances in neural information processing systems.* 2014. 3104-3112.

James Martens، و Ilya Sutskever. "Learning recurrent neural networks with hessian-free optimization." *Proceedings of the 28th International Conference on Machine Learning (ICML-11).* 2011. 1033-1040.

Jan Chorowski، Dzmitry Bahdanau، Kyunghyun Cho، و Yoshua Bengio. "End-to-end continuous speech recognition using attention-based recurrent nn: First results." *arXiv preprint arXiv:1412.1602*، 2014.

Jan Chorowski، Dzmitry Bahdanau، Kyunghyun Cho، و Yoshua Bengio. "End-to-end continuous speech recognition using attention-based recurrent nn: First results." *arXiv preprint arXiv:1412.1602*، 2014.

Jeffrey L. Elman. "Finding structure in time." *Cognitive science* (Wiley Online Library) 14 (1990): 179-211.

Jeffrey L. Elman. "Finding structure in time." *Cognitive science* (Wiley Online Library) 14 (1990): 179-211.

John Duchi، Elad Hazan، و Yoram Singer. "Adaptive subgradient methods for online learning and stochastic optimization." *Journal of Machine Learning Research* 12 (2011): 2121-2159.

Jonathan Sloculn. "A SURVEY OF MACHINE TRANSLATION: ITS HISTORY, CURRENT STATUS, AND FUTURE PROSPECTS ." *Computational Linguistics*، 1985، الإصدار 11.

Joshua T. Goodman. "A bit of progress in language modeling." *Computer Speech & Language* (Elsevier) 15 (2001): 403-434.

Kelvin Xu، وآخرون. "Show, attend and tell: Neural image caption generation with visual attention." *International conference on machine learning.* 2015. 2048-2057.

Kenneth Heafield. "KenLM: Faster and smaller language model queries." *Proceedings of the sixth workshop on statistical machine translation.* 2011. 187-197.

Klaus Greff، Rupesh K. Srivastava، Jan Koutnı́k، Bas R. Steunebrink، و Jürgen Schmidhuber. "LSTM: A search space odyssey." *IEEE transactions on neural networks and learning systems* (IEEE) 28 (2017): 2222-2232.

Kyunghyun Cho، وآخرون. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." *arXiv preprint arXiv:1406.1078*، 2014.

Libin Shen، Jinxi Xu، و Ralph Weischedel. "A new string-to-dependency machine translation algorithm with a target dependency language model." *Proceedings of ACL-08: HLT*، 2008: 577-585.

Luong Minh-Thang، Sutskever Ilya، V. Le Quoc، Vinyals Oriol، و Zaremba Wojciech. " Addressing the rare word problem in neural machine translation." *ACL*، 2015.

M Boualem. "CFP – JEP – TALN – Special Session on Arabic Language Processing." 2003.

Michel Galley، و Christopher D. Manning. "A simple and effective hierarchical phrase reordering model." *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* 2008. 848-856.

Mohammed N. Al، Taghreed M. Hailat، و Emad M. Al. "Evaluating English to Arabic Machine Translation Using BLEU." *(IJACSA) International Journal of Advanced Computer Science and Applications*، 2013، الإصدار 1.

Muriel Vasconcellos. "The Georgetown project and Leon Dostert: recollections of a young assistant, Early Years in Machine Translation." *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 3* (JOHN BENJAMINS PUBLISHING COMPANY)، 2000: 87-96.

Nal Kalchbrenner، و Phil Blunsom. "Recurrent continuous translation models." *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing.* 2013. 1700-1709.

—. "Recurrent continuous translation models." *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing.* 2013. 1700-1709.

Pascanu Razvan، Mikolov Tomáš، و Bengio Yoshua. "On the difﬁculty of training recurrent neural networks." *ICML*، 2013.

Paul J. Werbos، و others. "Backpropagation through time: what it does and how to do it." *Proceedings of the IEEE* 78 (1990): 1550-1560.

Peter F. Brown، Vincent J. Della Pietra، Stephen A. Della Pietra، و Robert L. Mercer. "The mathematics of statistical machine translation: Parameter estimation." *Computational linguistics* (MIT Press) 19 (1993): 263-311.

Peter Sheridan. "Research in language translation on the IBM type 701." *IBM Technical Newsletter* 9 (1955): 5-24.

Philipp Koehn، Franz Josef Och، و Daniel Marcu. "Statistical phrase-based translation." *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1.* 2003. 48-54.

Rafal Jozefowicz، Oriol Vinyals، Mike Schuster، Noam Shazeer، و Yonghui Wu. "Exploring the limits of language modeling." *arXiv preprint arXiv:1602.02410*، 2016.

Razvan Pascanu، Tomas Mikolov، و Yoshua Bengio. "On the difficulty of training recurrent neural networks." *International conference on machine learning.* 2013. 1310-1318.

Roland Kuhn، و Renato De Mori. "A cache-based natural language model for speech recognition." *IEEE transactions on pattern analysis and machine intelligence* (IEEE) 12 (1990): 570-583.

Roni Rosenfeld. "A maximum entropy approach to adaptive statistical language modeling." *Computer Speech and Language*، 1996: 187-228.

Salah El Hihi، و Yoshua Bengio. "Hierarchical recurrent neural networks for long-term dependencies." *Advances in neural information processing systems.* 1996. 493-499.

Sepp Hochreiter، و Jürgen Schmidhuber. "Long short-term memory." *Neural computation* (MIT Press) 9 (1997): 1735-1780.

Spence Green، Sida Wang، Daniel Cer، و Christopher D. Manning. "Fast and adaptive online training of feature-rich translation models." *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* 2013. 311-321.

Stanley F. Chen and Joshua Goodman. "An empirical study of smoothing techniques for language modeling." *34th Annual Meeting of the Association for Computational Linguistics (ACL).* 1996. 310-318.

Stanley F. Chen. "Shrinking exponential language models." *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics.* 2009. 468-476.

Stanley F. Chen، و Joshua Goodman. "An empirical study of smoothing techniques for language modeling." *Computer Speech & Language* (Elsevier) 13 (1999): 359-394.

Stanley F. Chen، و Ronald Rosenfeld. "A survey of smoothing techniques for ME models." *IEEE transactions on Speech and Audio Processing* (IEEE) 8 (2000): 37-50.

Thorsten Brants، Ashok C. Popat، Peng Xu، Franz J. Och، و Jeffrey Dean. "Large language models in machine translation." *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL).* 2007.

Toma´sˇ Mikolov. " Statistical Language Models Based on Neural Networks." Ph.D. dissertation، Brno University of Technology.، 2012.

Tomáš Mikolov، Martin Karafiát، Lukáš Burget، Jan Černockỳ، و Sanjeev Khudanpur. "Recurrent neural network based language model." *Eleventh annual conference of the international speech communication association.* 2010.

Tomáš Mikolov، Stefan Kombrink، Lukáš Burget، Jan Černockỳ، و Sanjeev Khudanpur. "Extensions of recurrent neural network language model." *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 2011. 5528-5531.

Tomas Mikolov، و Geoffrey Zweig. "Context dependent recurrent neural network language model." *2012 IEEE Spoken Language Technology Workshop (SLT).* 2012. 234-239.

Tsungnan Lin، Bill G. Horne، Peter Tino، و C. Lee Giles. "Learning long-term dependencies in NARX recurrent neural networks." *IEEE Transactions on Neural Networks* (IEEE) 7 (1996): 1329-1338.

Volodymyr Mnih، Nicolas Heess، Alex Graves، و others. "Recurrent models of visual attention." *Advances in neural information processing systems.* 2014. 2204-2212.

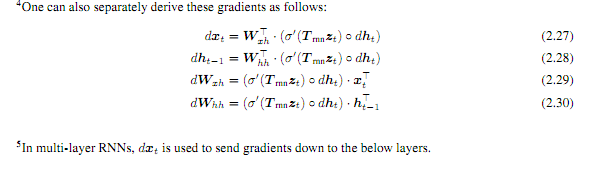
W. John Hutchins. "Machine translation: a concise history." *Computer Aided Translation: Theory and Practice*، 2007.

Warren Weaver. "Translation." تأليف *Machine Translation of Languages*، تحرير William N. Locke، و A. Donald Boothe، 15-23. Cambridge, MA: MIT Press، 1949/1955.

Wojciech Zaremba، Ilya Sutskever، و Oriol Vinyals. "Recurrent neural network regularization." *arXiv preprint arXiv:1409.2329*، 2014.

Yoshua Bengio، Nicolas Boulanger-Lewandowski، و Razvan Pascanu. "Advances in optimizing recurrent networks." *2013 IEEE International Conference on Acoustics, Speech and Signal Processing.* 2013. 8624-8628.

Yoshua Bengio، Patrice Simard، Paolo Frasconi، و others. "Learning long-term dependencies with gradient descent is difficult." *IEEE transactions on neural networks* 5 (1994): 157-166.

1. Question: If V is the size of the target vocabulary, how many are there for a sentence of length T? [↑](#footnote-ref-2)
2. Question: How many parameters does an n-gram model with a particular n have? [↑](#footnote-ref-3)
3. بعض المراجع تسمي هذه النماذج maximum entropy language model [↑](#footnote-ref-4)
4. [↑](#footnote-ref-6)
5.  [↑](#footnote-ref-7)
6. 2

   Eq. (4.4) implies that all alignment vectors at are of the same length. For short sentences, I only use the

   top part of at and for long sentences, I ignore words near the end. [↑](#footnote-ref-8)
7. I will refer to this difference again in Section 4.1.3. [↑](#footnote-ref-9)
8. If the window crosses the sentence boundaries, I simply ignore the outside part and consider words inthe window [↑](#footnote-ref-10)
9. local-m is the same as the global model except that the vector at is ﬁxed-length and shorter. [↑](#footnote-ref-11)
10. local-p is similar to the local-mmodel except that I dynamically compute pt and use a truncatedGaussian

    distribution to modify the original alignment weights align(ht, ¯ hs) as shown in Eq. (4.6). By utilizing pt to

    derive at, I can compute backprop gradients forWp and vp. This model is differentiable almost everywhere. [↑](#footnote-ref-12)