片語式機器翻譯中未知詞與落單字的問題探討

*蔣明撰　+黃仲淇　*顏合淨　*黃士庭　*張俊盛　++楊秉哲　++谷圳
*國立清華大學資訊工程學系
+國立清華大學資訊系統與應用研究所
++資訊工業策進會
{raconquer, u901571, fi26.tw, koromiko1104, jason.jschang}@gmail.com
++{maciaclark, cujing}@iii.org.tw

摘要

近年來，機器翻譯技術蓬勃發展並越顯重要。然而，現存的機器翻譯系統對於（系統未收錄）未知詞多採直接輸出到目標翻譯的方式。此忽略的舉動可能造成未知詞附近的選字錯誤，或是其附近的翻譯字詞順序錯置，因而降低翻譯品質或降低閱讀者對翻譯文章的理解。經過我們的初步分析，大約有 25%的系統未知詞可用重述（paraphrase）的方式來作翻譯，另外的 25%可利用組合單字翻譯來翻譯。另外，現有的片語式（phrase-based）機器翻譯系統對於落單字（singleton）的翻譯效果也未加重視。所謂的落單字是指系統在翻譯此字時必須單獨翻譯：此字沒法與前面或是後面的字組合成連續字詞片語或是文法翻譯結構。本研究將建構於片語式機器翻譯處理技術，開發未知詞翻譯模組和落單字翻譯模組。實驗結果顯示即使在不假額外的雙語資料，我們的未知詞翻譯模組仍勝出片語式翻譯系統，尤其是在包含有未知詞的句子上。

關鍵詞：未知詞，重述，片語式機器翻譯系統，落單字，機器翻譯

一、緒論

近年來，機器翻譯技術蓬勃發展並越顯重要。然而，現今先進的片語式機器翻譯系統對於（系統未收錄）未知詞與落單字（singleton）的處理仍有改進的空間。翻譯系統對於來源語（source language）未知詞採直接輸出到目標（target language）翻譯的方式，也就是說，系統並不處理未知詞。此忽略的舉動可能造成未知詞附近的選字錯誤，或是其附近的翻譯字詞順序錯置，因而降低翻譯品質或降低閱讀者對翻譯文章的理解。片語式機器翻譯系統之所以可以有令人滿意的翻譯效果在於其翻譯的過程常常是多個連續的來源語字詞一起翻譯到目標語。多個字詞一起翻譯的過程幫助了這些字詞翻譯的解歧，也就是所謂的字義解歧（Word Sense Disambiguation）亦或是字詞翻譯解歧（Word Translation Disambiguation）。以中文字「起」爲例。「起」有相當多的字義如「起床」、「上升」、「動身」、「發揮」等。不同字義的（英文）翻譯也都不盡相同。而片語式翻譯系統則會將「起」跟其周遭連續的字「的」、「很」和「早」一起看作是一個片語並翻譯成“get up very early”。換言之，解歧成「起床」字義。很少文獻針對片語式機器翻譯系統中的落單字翻譯效果進行分析。所謂的落單字是指系統在翻譯此字時必須單獨翻譯：此字無法與前面或是後面的字組合成連續字詞片語或是文法翻譯結構。落單字必然是片語式翻譯系統的自然天敵。目前系統多靠語言模型（Language Model）來選擇落單字的翻譯。但是語

言模型受限於字數限制，也不考慮像是字詞詞性等語言現象，大多數選擇最高頻的翻譯。落單字的翻譯解歧效果直接影響了翻譯之品質。

首先，我們分析了 NIST MT-08 的測試句。美國 NIST（National Institute of Standards and Technology）幾乎每年都會舉辦 MT 的比賽來促進自動翻譯研究的發展。經過我們的初步分析，大約有 25%的系統未知詞可用重述（paraphrase）的方式來作翻譯，另外的 25%可利用組合單字翻譯來翻譯。重述就是將未知詞轉換成意思相近但現於現有雙語語料中的字詞。重述的論文探討已經相當多且齊全。在這個計畫中，我們將著重在跟重述佔有相同重要角色的組合單字翻譯上。我們利用組合單字的翻譯來翻譯未知詞。我們的處理方法不假額外的雙語資料（文獻多直接藉由擴大雙語語料來減少未知詞），只利用現存的訓練資料來尋找可能的單字翻譯，也就是，系統已知字詞（in-vocabulary）翻譯。更精確的來說，我們組合排列現有的雙語訓練資料中未知詞的構成字之翻譯，並加以排序以得到較爲可能的未知詞翻譯。例如：藉由雙語資料中「上」的翻譯 upper、above、rise 等，以及「肢」的翻譯 body、limbs 等可組合出 NIST MT-06 未知詞「上肢」的翻譯 upper limbs。類似的方法可以組合出形容詞-名詞複合字未知詞「韓戰」（Korean war），名詞-名詞複合字未知詞「邊貿」（border trade），動詞-形容詞複合字未知詞「成名」（become famous）之翻譯。其中，「邊貿」也是目前最尖端的翻譯系統 Google Translate 之未知詞。

另外，在針對片語式機器翻譯中落單字的翻譯時，我們發現，隨機抽樣 NIST MT-08 的五十中文句中，落單字佔全文比例高於 6%，落單字又以名詞、動詞居多，各佔 72%、21%。人工分析系統對於不同詞性字詞的翻譯品質差異很大，名詞可達五成正確率（precision），但是動詞只到兩成。分析 NTCIR 2011 年專利翻譯比賽的發展中資料，也顯示了類似比例—落單字佔全文比例約 5%。由上面幾組數據，我們知道落單字跟未知詞一樣，都是片語式機器翻譯系統急須面對處理的課題。我們預計利用「動詞-名詞」或是「動詞-副詞」搭配詞（collocation）來幫助落單字的解歧，以增加片語式機器翻譯系統之翻譯品質。畢竟，落單字要解歧就需要看稍微遠一點的字詞（context），而搭配詞往往又是幫助解歧的有用資訊（一個搭配詞一個字義 one sense per collocation）。以「起」和「打擊」這兩個多義字來作說明。它們的翻譯可能爲 get up、rise、increasing、play、have 等，和 fight、combat、batting、bat 等。但是當「起」的附近有搭配詞「早」時 get up 較有可能，當附近有名詞搭配詞「作用」時 play、have 較有可能（此時的「起」有「發揮」的意思）。類似地，當「打擊」附近有搭配詞「犯罪」出現時 fight、combat 較有可能，而當其附近出現「區」，「棒球」時，則是 batting、bat 較爲可能。由上面的例子，我們預期：不一定緊密相鄰的「動詞-名詞」或是「動詞-副詞」搭配詞，或稱爲有彈性的搭配詞（flexible collocation），將可幫助片語式機器翻譯中落單字解歧。

本研究將建構於現有片語式機器翻譯處理技術，例如公開原始碼的 Moses 翻譯系統，開發未知詞翻譯模組和落單字翻譯模組。未知詞翻譯模組將從現存的雙語訓練資料中尋找未知詞構成字之翻譯，進而組合、排序未知詞的翻譯候選（利用雙語對應機率和單語流暢度加以排序）。排序好的翻譯候選將利用 XML 標記方法輸入片語式機器翻譯系統以作句子翻譯。落單字翻譯模組則會先利用大量的中文語料（如：Chinese Gigawords）抽取出數學統計上可能的搭配詞如「起…作用」、「打擊…犯罪」等。然後藉由這些搭配詞來爲落單字解歧。解歧完後的落單字翻譯也是利用 XML 標記方法將翻譯候選提供給真正作句子翻譯的片語式機器翻譯系統。所以我們的方法除了使用雙語資料外，也會利用中文語料與英文語料（如： English Gigawords）取得中文搭配詞和英文語言模型。

二、研究方法（The Method）

本研究的範圍在於解決一般機器翻譯最常忽略的未知詞翻譯問題還有落單字的翻譯解歧問題。目標是，在現有雙語訓練語料中，爲未知詞找出翻譯並有效排序翻譯候選，另外，正確替落單字解歧，提升機器翻譯品質亦或是幫助閱讀者閱讀。我們將在以下章節詳述建構在現有片語式機器翻譯系統之上的未知詞翻譯模組和落單字解歧模組。

（一） 未知詞翻譯模組

未知詞翻譯模組針對未收錄於機器翻譯訓練語料的字詞產生並依照可能機率排列其翻譯候選。此模組可分爲兩個子模組—組成字模組和重述模組（目前我們較著重在文獻較少提到的組成字模組）。

1. 組成字模組

未知詞是系統未收錄的字詞，也就是，利用完全無誤比對（exact-match）來查詢雙語語料以得目標語翻譯必定是徒勞無功的。此模組將原本完全無誤比對（exact-match）的翻譯查詢轉換成一連串的部分比對（partial-match）查詢以先求得未知詞構成字的翻譯。接著從這些查詢回來的雙語配對（phrase pair）中，擷取出未知詞組成字的可能翻譯。最後，藉由組合組成字翻譯來翻譯未知詞，並且參考雙語字層級（character-level）翻譯機率和目標語的語言模型來排序未知詞翻譯候選。步驟大綱如下。

步驟一：我們將原本毫無所獲的字詞翻譯查詢轉換成一系列的萬用字元（wildcard）查詢以得組成字之可能翻譯。舉例來說，在不增加或是改變雙語語料的情況下，我們將對於未知詞「上肢」毫無斬獲的完全比對翻譯查詢變成「上*」和「*肢」的部分比對系列查詢，可查到翻譯配對如（上訴, appeal for）、（上升, increasing of）、（上段, upper block）等，和（四肢, the body）、（四肢, four limbs）、（義肢, prosthesis）等。

步驟二：上一個步驟得到的是來源語的字詞翻譯而不是未知詞組成字的翻譯，也就是，不是字層級（character-level）的翻譯。所以此步驟首先擷取出組成字的翻譯可能。我們是利用 N-gram 來擷取出組成字的可能翻譯。以翻譯配對（上段, upper block）和（四肢, four limbs）爲例。未知詞的組成字「上」和「肢」的可能翻譯分別是“upper”、“block”、“upper block” 和 “four”、“limbs”、“four limbs”。值得注意的是，產生 N-gram 時，我們會考慮其變化型。這些產生的 N-gram，其實詞（例如名詞、動詞等）限定必須出現在一個大的字詞語料庫中（例如 WordNet），如果沒被此大的語料庫所包含將被剔除：畢竟一個沒被字詞語料庫包含的實詞，其 N-gram 應該也不是怎樣好的翻譯候選。最後，我們排除低頻的 N-gram。爲了公平的比較，次數是變化型的累加並共享。爲了得到原形化資訊，我們實作時，利用 NLTK 中提供的原形化器（Bird 等人, 2008）。表一呈現步驟一和步驟二的個別產物。

表一：步驟一和步驟二的輸出產物

步驟一		步驟二	
來源字詞	目標 phrase	來源字詞	目標 N-grams
四肢	the body	四肢	body
四肢	extremities	四肢	extremity
		四肢	extremities
四肢	four limbs	四肢	four
		四肢	limb
		四肢	limbs
		四肢	four limbs
義肢	prosthesis	義肢	prosthesis

表二：所有組成方法和特色組成方法所產生的雙語關聯例子。

字典中配對		雙語關聯	
source phrase	*translation*	所有組成方法 All Constituent	特色組成方法 Salient Constituent
肢	limb	(肢, limb)	(肢, limb)
手足	limb	(手, limb)	(足, limb)
		(足, limb)	
肢體	limb	(肢, limb)	(肢, limb)
		(體, limb)	
後肢	hind limb	(後, hind)	(肢, hind)
		(肢, hind)	(肢, limb)
		(後, limb)	(肢, hind limb)
		(肢, limb)	
		(後, hind limb)	
		(肢, hind limb)	

步驟三：我們利用雙語對應關係來刪除較不可能的組成字翻譯。步驟二所產生的 N-gram 有時候跟組成字的關聯是相當相當少的。爲了減少計算量和增加翻譯的準確度，我們將去除比較不可能的組成字翻譯 N-gram。以針對部分比對查詢「*肢」所找出來的翻譯配對（四肢, four limbs）爲例。因爲“four”和“limbs”皆是常見且高頻的實詞，步驟二將會保留兩者，並視爲組成字「肢」的可能翻譯。我們很明顯的知道雖然“limbs”是其合理的翻譯，但是中文應該是「四」才對的“four”顯然不是。也因此需要此步驟來檢驗存留下來的組成字翻譯和組成字的關係強弱。

首先，我們利用雙語字典如 bilingual WordNet 來建立雙語對應關係。建立關係的方

式可分爲兩種方法—所有組成和特色組成方法。我們詳述如下。

- 所有組成（all-constituent）方法：針對每一個字典中的翻譯配對<*source phrase*, *translation*>，我們爲 *source phrase* 中的所有組成字和 *translation* 中的所有 N-gram 建立起對應關係。也就是說，一旦一個 *source phrase* 中的組成字和 *translation* 中的 N-gram 有共同出現過，他們之間就會有一個連結。以字典中的<“後肢”, “hind limb”>爲例。「後」和「肢」這兩個構成字將會和“hind limb”的 N-gram 有所連結。我們會爲此配對建立 6 個雙語關聯（請參見表二）。

- 特色組成（salient-constituent）方法：相較於上述方法，此方法只會爲 *source phrase* 中的特色組成字和 *translation* 的 N-gram 建立關聯。一個 *source phrase* 中的組成字是特色組成字如果此組成字和 *translation* 是最有相關的。嚴謹的來說，針對字典中配對<*source phrase*, *translation*>，特色組成字 *c**是利用下面的公式而得

$$\arg\max_{c} Dice(c, translation) = \arg\max_{c} \frac{2 \cdot Count(c, transaltion)}{Count(c) + Count(translation)}$$

其中 *c* 代表 *source phrase* 中的組成字，而 *Count*（·）代表字典內的頻率。以<“後肢”, “hind limb”>爲例。我們比較 *Dice*(“後”, “hind limb”) 和 *Dice*(“肢”, “hind limb”)來決定特色組成字。因爲 *Count*(“後”, “hind limb”)和 *Count*(“肢”, “hind limb”)爲 1 且「後」、「肢」、和“hind limb”發生次數個別爲 1073、201、1，因此擁有較高 Dice 值的「肢」被選爲「後肢」的特色組成字，進而註冊雙語關聯(“肢”, “hind”)、(“肢”, “limb”)、和(“肢”, “hind limb”)（可參見表二）。我們可以知道特色組成方法所產生的雙語關聯將是所有組成方法所產生的子集合。

一旦建立起字典的雙語關聯，我們將可以刪除沒出現在關聯內的組成字和其上一步驟產生的 N-gram 配對。舉例來說，針對組成字「肢」所找到的翻譯配對（四肢, four limbs）及步驟二所允許的 N-gram（四肢, four）將在步驟三中被去除。因爲組成字和其 N-gram 配對（肢, four）沒在表二中出現。在實作上，我們首先利用所有組成方法來去除翻譯候選且保存高召回率。如果存留下來的組成字翻譯候選仍是超過門檻值（threshold），我們再使用特色組成方法來更積極作刪除以達到高準度。另外，這些雙語關聯也用作軟限制（soft constraint）而其頻率則當成是下一步驟排序的特徵（feature）。

步驟四：我們利用圖一的演算法來組合出並排列未知詞的翻譯候選。首先我們爲未知詞 *O* 的每一個組成字 *c* 從雙語翻譯對應 *TE* 中抽取出其翻譯 *SubTrans*（利用上述步驟一到三）。*SubTrans* 是一個 list 其元素像（*source word*, *target N-gram*），其中 *source word* 包含了 *O* 的組成字。然後（圖一步驟 1b），我們使用雙向條件機率（bidirectional conditional probabilities）來測量組成字和其翻譯的雙語關聯度，並將這樣的資訊紀錄在相對應的字層級（character-level）位置上。*CandList* 內的元素將像(*c*,（*source word*, *target N-gram*）, P(*target N-gram*|*c*)· P (*c*|*target N-gram*))。其中雙向條件機率 P(*target N-gram*|*c*)和 P(*c*|*target N-gram*)則是由字層級（character-level）對應的平行語料（parallel corpus）訓練而來。以未知詞「上肢」爲例。我們首先爲組成字「上」和「肢」取得 *SubTrans*{(“上訴”, “appeal”), (“上策”, “policy”), …, (“上段”, “upper”)}和{(“四肢”, “limb”), (“四肢”, “limbs”), …, (“義肢”, “prosthesis”)}。然後我們計算組成字和其 N-gram 的對應強度並將這些資訊紀錄在 *CandList* 中（可參考表三）。

```
procedure GenerateAndEvaluateCandidates(O, TE, C, CT)
    for each constituent c in the OOV O
(1a)    SubTrans = RetrieveSublexicalTranslations(c, O, TE)
(1b)    CandList[position (c, O)] = BilingualInfo(SubTrans, c, C)
(2a) Straight = CandList[1]
(2b) Inverted = CandList[|O|]          // where |O| denotes the length of O
    for each constituent position cp >1 in ascending constituent positions of O
(3a)    Straight⊗ = CandList[cp]
    for each constituent position cp <|O| in descending constituent positions of O
(3b)    Inverted⊗ = CandList[cp]
(4a) Straight = MonolingualInfo(Straight, CT)
(4b) Inverted = MonolingualInfo(Inverted, CT)
    Candidates = Straight + Inverted
(5)  RankedCandidates = Sort Candidates in decreasing order of probability P
(6)  Return the top N RankedCandidates with probabilities P exceeding θ
```

圖一：組合並排列未知詞之翻譯候選

表三：針對未知詞「上肢」的 *CandList* 樣本

CandList	*c*	*source word*	*target N-gram*	P(*target N-gram*|*c*) · P (*c*|*target N-gram*)
CandList[1]	上	上訴	appeal	5×10^{-5} · 0.17
	上	上策	policy	1.2×10^{-7} · 1×10^{-9}
	上	上段	upper	0.02 · 0.56
CandList[2]	肢	四肢	limb	0.05 · 0.01
	肢	四肢	limbs	0.05 · 0.01
	肢	義肢	prosthesis	0.004 · 0.12

一旦我們有組成字的翻譯，我們便可產生未知詞翻譯候選。雖然未知詞的翻譯範圍遠小於翻譯一整個句子。翻譯的重組（re-ordering）仍是有可能發生。例如，「調」和「氣」的個別翻譯是“adjustment”和“air”，「調氣」的翻譯則是倒置成“air adjustment”。也因此，全順接（straight）和全反接（inverted）的情況都會被考慮。在圖一步驟 3 中，Straight 和 Inverted 會接續的涵蓋未知詞的組成字：邊收集組成字翻譯邊累乘翻譯的機率。每一個組合而成的翻譯候選 *TransCand* 的字詞翻譯分數是由雙向條件機率的乘績來推估。計算方式如下：

$$P_{trans} = \sqrt[|o|]{\prod_{c_i \in o} p(c_i \mid target\ N-gram_{ij}) \cdot P(target\ N-gram_{ij} \mid c_i)}$$

其中 c_i 代表未知詞的組成字而 *target N-gram*$_{ij}$ 代表 c_i 其中一個組成 *TransCand* 的翻譯。以未知詞「上肢」之組成字翻譯(“上”, (“上段”, “upper”), 0.02 · 0.56)和(“肢”, (“四肢”, “limb”), 0.05 · 0.01)爲例。我們會產生一個全順接的翻譯候選(“上肢”, “upper limb”, $((0.02\cdot 0.56)(0.05\cdot 0.01))^{\frac{1}{2}}$)和一個全反接的(“上肢”, “limb upper”, $((0.02\cdot 0.56)(0.05\cdot 0.01))^{\frac{1}{2}}$)。

除了利用雙語資訊外，我們也利用了單語資訊來檢測翻譯候選。每一個翻譯候選的 Mutual Information（*MI*）值將會利用下式計算出來。

$$MI(w_1, w_2) = \log_2\left(\frac{\Pr(w_1, w_2)}{\Pr(w_1)\Pr(w_2)}\right)$$

其中 w_1 和 w_2 是 *TransCand* 中的 bigram。對於 *MI* 值超過門檻值的翻譯候選我們將會算出目標語言的語言模型機率 $P_{TLM}(TransCand)$並將其乘在字詞翻譯機率上如下式（以得到評量翻譯候選的分數）

$$Score(TransCand) = P_{trans}(TransCand)^{\lambda_1} \cdot P_{TLM}(TransCand)^{\lambda_2}$$

其中 λ_i 是特徵權重值而 $\sum\lambda_i$ 等於 1。P_{TLM} 用來幫助辨識組合的翻譯候選的流暢度（fluency）。

演算法最後回傳前 N 個 *Score* 值超過門檻值 θ 的翻譯候選。這些候選將被當作是可能的未知詞翻譯。發動未知詞模組的門檻值 θ 和 N 將會利用發展中資料來尋找。

2. 重述模組

重述模組的目標在於將系統未收錄的未知詞轉換成意思相近或同義的系統已知詞（in-vocabulary）。再藉由已知詞來取得對應未知詞的翻譯。以未知詞「中餐」爲例。我們首先將其轉換成翻譯系統收錄並相似或同義的詞「午餐」、「午飯」等，再排序這些詞的翻譯當作是「中餐」的翻譯候選。

我們可以利用手工編撰辭典（thesaurus）或是機器學習（machine learning）技術來重述未知詞。手工編撰的資源如同義詞詞林或是 Sinica BOW （Bilingual Ontology WordNet）可加以利用。同義詞詞林可以直接提供高準度的未知詞同義字，而 Sinica BOW 則可利用翻譯相同或是近似來提供未知詞的重述。利用翻譯的重述，文獻上將其稱爲依賴第二語言（此例爲英文）的字詞層級（lexical-level）重述。舉例來說。在一個雙語字典或是漢英字典裡面，如果收錄翻譯條目（ “中餐”, “lunch” ）和（ “午餐”, “lunch” ），我們將可以知道「中餐」和「午餐」同屬一個翻譯，在某些情境下同義，可互換，互爲重述。仰賴第二語言的重述，其來源不一定是高準度的人工編撰辭典，也可能是高涵蓋率的自動字詞對應（word alignment）結果（其訓練資料通常是做好句子對應的平行語料，如 Marton 等人），亦或是混合高準度的字典和高涵蓋率的自動對應。

機器學習技術如分布相似法（distributional similarity）或是最大熵值法（maximum entropy）可推敲習得並排序未知詞的可能相似詞或是同義字。詳細地說，我們利用大量

中文語料（如 Chinese Gigawords）中各字詞的前後文字（context words）來分析哪些文字附近的字詞非常接近。前後文字接近的字詞則可以視爲互爲重述。例如：互爲重述的兩詞「中餐」和「午餐」前後皆常出現「吃」和「享用」。在實作面上，模組可將重述詞限制在高頻且出現雙語語料中的數千個中文詞，以減少計算量。另外，爲了避免資料稀疏（data sparseness）問題，也許可以考慮字詞的類別而不是字的本身。手工辭典包含分類訊息如 Sinica BOW 或 E-HowNet 或是利用單語或雙語之自動字詞分類技術如（Och 1999），將可提供類別資訊。

相較於上述互爲對等的重述，Mirkin 等人在 2009 年利用推演規則（entailment rule）來重述未知詞。所謂的推演規則是：如果 A 可以推演出 B，那麼 B 就是 A 較爲一般的詞語。也就是說 A 是比較特定（specific）的詞，而 B 則比較一般（general）。舉例來說。未知詞“skyscraper”可以推演出“building”，那麼我們就可以利用“building”的翻譯來翻譯“skyscraper”。Mirkin 等人運用 WordNet 的上、下位詞資訊來取得推演規則（其重述的過程仍是有包含對等重述，利用的是 WordNet 的同義詞群組（synset））。實驗結果顯示「重述後的未知詞，有助於產生讀者更易了解原文的翻譯，潛在地，對於後處理編輯（post-editing）有所幫助」。

（二） 落單字翻譯模組

當一個單字無法與週遭字詞結合成連續片段時，我們說此單字落單存在，無法型成片語。對於這樣的落單字，片語式機器翻譯系統很難利用它前後字詞搭配解歧的優勢，爲其找出適當正確的翻譯。此模組架構在現有的片語式機器翻譯系統之上，以預處理（pre-processing）的方式爲落單字解歧（減少、限定翻譯候選），以避免片語式系統以受限的語言模型單獨翻譯落單字。

(1) 藉由單語語料取得數學統計上可信的搭配詞
(2) 從翻譯表中推衍出跳躍式 bigram 配對（skipped bigram pair）
(3) 輸出上述產物

圖二：解歧階段之前處理

圖二顯示該階段之前處理步驟。我們首先利用（Smadja, 1993）的方法從大量的單語語料庫（如：Chinese Gigawords）中抽取搭配詞。Smadja 的方法留意兩個單字是否常常一起出現、是否在某種距離下常常一起出現。舉例來說。從語料庫中我們可以發現「起」、「作用」常常距離一，因爲夾著形容詞（如：「起 極大 作用」、「起 承先啓後 作用」），也常常距離二，因爲多了副詞「了」（如：「起 了 極大 作用」、「起 了 正面的 作用」）。經過 Smadja 方法中的 *MI* 值、距離標準差、還有變異數的過濾篩選後，我們將所留下來的字詞搭配視爲數學統計上可信的搭配詞。

有了單語搭配詞後，我們還需要搭配詞的翻譯。在不增加雙語訓練資料的前提下，我們利用底層（underlying）片語式機器翻譯系統如 Moses 從平行語料中產生的翻譯表（phrase table）來推演出搭配詞翻譯。Moses 首先利用 GIZA++來作字詞對應（word alignment），接下來套用 grow-diag-final 的演算法來合併 GIZA++雙向的字詞對應結果。這樣產生的翻譯內容如集合 {（“起 正面 作用”, “play a positive role”）, （“起 正面 作用”, “have a positive effect”）, …, （“早 起”, “get up early”）, （“起 得 很

早”，“get up very early”），…}。

爲了解歧落單字—它與前後字合成的片段未見於翻譯表中，我們將在翻譯表的翻譯配對中跳躍的選取來源語的字詞，並利用翻譯表中字詞對應結果來選擇這些字詞在目標語的翻譯。 以翻譯配對（ “起 正面 作用”，“play a positive role”）和（ “起 得 很 早”，“get up very early”）爲例。注意：字詞配對以反白還有下底線來呈現。如果我們跳躍的選擇「起」和「作用」，我們可以得到翻譯配對（ “起 … 作用”，“play … role”）；如果我們跳躍的選擇「起」和「早」，我們可以得到翻譯配對（ “起 … 早”，“get up … early”）。我們將這樣跳躍取得的翻譯配對稱作跳躍式 bigram 配對，因爲在中文端限制兩個字詞。

在執行時，我們首先在句子的固定範圍（window size）內尋找落單字的搭配詞。利用較爲可能的搭配詞來限定落單字的翻譯候選。例如在句子「我國警方有效打擊青少年國際犯罪」的落單字「打擊」有其名詞搭配詞「警方」、「犯罪」，形容詞搭配「國際」。利用這些搭配詞的對應翻譯如（ “警方…打擊”，“the police … fight”）、（ “打擊…犯罪”，“fight … crimes”）可抽取出落單字的翻譯候選 fight。以降低或直接排除「打擊」在該上下文中較不可能的翻譯如 bat 和 batting 的排序。同樣地，「起」可以分別利用「早」和「作用」解歧成“get up”、“play”和 “have”。

三、實驗設定

目前實驗專注在未知詞組合字翻譯模組上，未來也會涵蓋其他如落單字解歧模組的實驗分析。在此章節中，我們首先介紹底層的（underlying）機器翻譯系統 Moses 和我們如何將翻譯候選加入此系統（章節 3.1）。再來，我們敘述實驗中會用到的資料，包含訓練還有發展資料。章節 3.3 則是敘述我們如何根據發展中資料來更改查詢方式以取得未知詞翻譯候選。最後，我們描述微調與設定系統參數的過程。

（一）底層片語式機器翻譯系統

我們所提出的未知詞組合字模組只針對未知詞提供翻譯候選，因此必須架構再現有的翻譯系統上。在實驗時，我們選用目前先進表現優異的片語式機器翻譯系統 Moses（Koehn 等人 2007）作爲我們的底層翻譯系統。Moses 提供簡單的 XML 標記語言讓外部模組所產生的單字或是字詞組的翻譯可以輕鬆被其利用，而不會更改到像是 Moses 內部的翻譯模組（translation model）與語言模組（language model）。

（二）資料集（data sets）

我們使用 Hong Kong Parallel Text（LDC2004T08）和 ISI 中英平行語料（LDC2007T09）來訓練 Moses 的翻譯模組（translation model）和重排模組（reordering model）。這些語料的中文部分是利用 CKIP 中研院斷詞器（Ma 和 Chen, 2003）來斷詞。我們使用標準化的設定來跑 Moses：跑 GIZA++（Och 和 Ney, 2003）來取得字詞對應、grow-diagonal-final 演算法（Koehn 等人, 2005）來結合雙向字詞對應結果、和在（Koehn 等人, 2005）內介紹的方法來抽取雙語對應。至於語言模型（language model），我們使用第三版 English Gigaword 中的新華新聞部份（LDC2007T07）。大約有 800 多萬個句子利用 SRILM 工具（Stolcke, 2002）來建立 trigram 的語言模型。

另一方面，我們的未知詞組合模組使用 WordNet 3.0（Miller 等人, 1990）和 Sinica BOW（Huang 等人, 2004）來過濾限制組成字的翻譯候選（章節 2.1.1）。在計算 *MI* 值上我們利用第三版 English Gigaword（LDC2007T07）和 Web 1T fivegram（LDC2006T13）資料。我們利用與訓練 Moses 相同的平行語料還有目標語語料來分別計算雙語組成字和目標語單字機率（也就是 bidirectional conditional probabilities）和目標語流暢度。

（三）查詢形式和雙語資源

表四：未知詞長度和個數分析（發展中資料）

未知詞長度	未知詞個數	百分比(%)
1	56	4.4
2	683	53.7
3	352	27.7
4	115	9
5+	67	5.3

我們利用 NIST MT-08 的資料來分析未知詞問題。在這份總共 1,357 句資料中，有 637 個句子有未知詞（共 1,273 個未知詞）。在這些未知詞中，我們將未知詞長度和個數關係列於表四。在後續實驗中，我們專注在幫助佔超過一半比例的雙字（two-character）未知詞尋找組合式翻譯候選。爲了更進一步分析未知詞的類型、查詢組合字翻譯的形式、適合查詢的雙語資料，我們隨機抽取 100 個包含有（至少一個）二字（two-character）未知詞的句子。表五成列出我們人工針對這 100 句中的未知詞所作的未知詞類型分析。我們在作分析時，會手動的將未知詞的翻譯從 NIST MT-08 的對應參考翻譯（reference translation）中標示出來。就像是圖六中包含有「上肢」未知詞的例句。我們的組合字翻譯模組是特別設計來處理表五中佔了四分之一強的 *combination forms* 未知詞。詳細資料可參考表五。

表五：未知詞型別、其定義和例子

未知詞型別	未知詞型別之定義	例子	未知詞個數
Order Variants	Sequence of characters reversed without changing the original meaning	療治(治療) (treat)	1
Writing Variants	Replacement between simplified and traditional Chinese characters	念書 (唸書) (study)	1
Domain Specific	Domain specific terminologies	勤務 (service support) 二傳 (setter)	2

Word + Suffix	Words composed by a content character (underscored character) and a not translated function character	忙著 (busy) 爐子 (stove)	4
Informal	Used in conversation or informal writing	看頭 (worth watching) 幹麼 (what)	6
Old Use	Words rarely in use now	古稀 (60 years old) 橫流 (all over)	8
Name Entity	Name entities could be transliterated such as person, place, and organization	布希 (bush) 膠州 (jiaozhou)	12
Segmentation Error	Words erroneously split by the segmentation system	領式 (開領式) 會兒 (這會兒)	16
Rare Paraphrase	Words could be translated by replacing with its paraphrases	踐行 (practice) 訪談 (interview)	25
Combination Form	Words could be translated by combining sublexical translations	上肢 (upper limbs) 肌力 (muscle strength)	25

直覺上，針對一個雙字未知詞 c_1c_2，有四種下萬用查詢的方式以得到組合字的翻譯。表六顯示第一種和第二種查詢形式可以找出最多翻譯候選。我們的模組就採用此兩種查詢形式。以未知詞「上肢」爲例。我們將會利用「上*」和「肢*」以及「上*」和「*肢」來查詢組合字的翻譯。

另一方面，在利用上述兩種查詢形式下，我們比較了不同雙語資料尋找組成字翻譯的有效度。我們比較了下面幾種資料的翻譯擊中率（translation hit rate）：林語堂的漢英字典（http://humanum.arts.cuhk.edu.hk/Lexis/Lindict/）、LDC 翻譯字典（LDC2002L27）、字層級的翻譯表（character-based phrase table）、和字詞層級的翻譯表（word-based phrase table）。在 25 個 *combination forms* 未知詞中，他們的擊中率分別是 0.64、0.68、0.60、0.88。字詞層級的翻譯表有最高的翻譯擊中率，因此被選爲我們查詢組成字翻譯的查詢對象。

表六：針對兩字詞 c_1c_2 使用不同查詢形式可翻譯的未知詞個數

查詢形式	可翻譯的未知詞個數	例子	
		未知詞	對應的中文字詞
「c_1*」和「c_2*」	17	上肢 (upper limbs)	(上方 肢體)
「c_1*」和「*c_2」	9	上肢 (upper limbs)	(上方 四肢)
「*c_1」和「c_2*」	2	震魔 (quake demon)	(地震 魔鬼)
「*c_1」和「*c_2」	1	鐘體 (bell body)	(時鐘 身體)

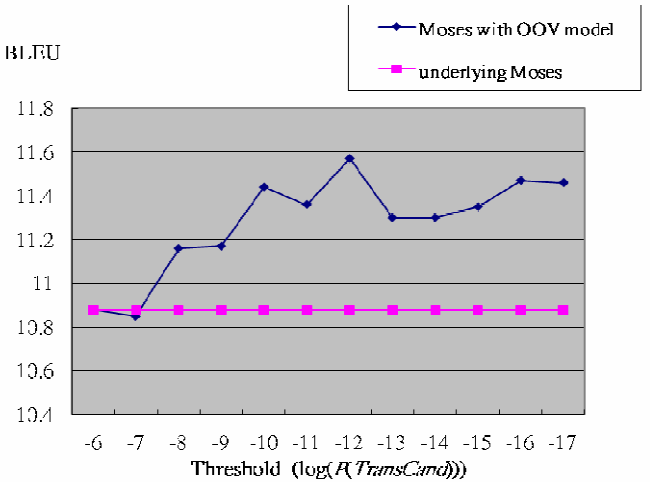
（四）參數設定

在這章節中，我們利用 50 句的發展資料來微調設定（tune）模組中的兩個參數—回傳的翻譯候選個數 *N* 還有被用來踢除較不可能的翻譯候選之門檻值 θ。這 50 句的發展資料每一句都有至少一個雙字未知詞，並且 25 句中的未知詞是型別 *combination forms*。

爲了選用一個適當的 *N*，我們首先觀察那 25 句包含有 combination-form 的未知詞對於不同 *N* 值的翻譯表現。在此，翻譯表現是由 *Mean Reciprocal Rank*（*MRR*）來作評估。*MRR* 被定義爲使用者在系統回傳的翻譯清單中定位第一個正確翻譯所需作的努力。*MRR* 介於 0、1 之間，1 又代表正確的翻譯總是在清單的最上頭。表七統整了不同 *N* 的涵蓋率和 *MRR* 值。在考量涵蓋率、*MRR*、和翻譯的時間複雜度（time complexity of decoding）之後，我們將 N 設爲 10。

表七：*N* 和 *MRR* 之關係表

N	在 25 個未知詞中可翻譯的個數	*MRR*
5	8	0.27
10	11	0.28
20	12	0.28
40	12	0.28



圖三：不同門檻值的 BLEU 翻譯表現

門檻值 θ 可以用來刪除候選也可以用來決定是否啓動我們的組合字翻譯模組因爲畢竟有些未知詞是不適合用這樣組合方式取得翻譯候選。較高的門檻值代表較少的翻譯候選，潛在地降低涵蓋率；而較低的門檻值代表較多的翻譯候選，潛在地降低準確度。爲了選用適當的 θ，我們將我們模組提供的未知詞翻譯候選用 XML 標記加入底層 Moses 系統，並且檢驗不同 θ 所得到的翻譯品質。在此，我們選用 BLEU（Papineni 等人, 2002）來當作翻譯品質的檢定標準。從圖三中可發現，當門檻值大於-8 時，相當少的翻譯候選會被考慮進去，導致翻譯品質並沒有差異太大；但是，當門檻值小於-13 時，有較多雜訊被考慮成翻譯，在翻譯的分數上就有所減少。我們選用擁有最好翻譯表現的-12 當作我們的過濾門檻值（大概在翻譯的準確度和涵蓋率取得平衡）。

四、評估

這一章節我們專注在評量未知詞組合字翻譯模組對片語式系統 Moses 帶來的影響。我們使用包含有 1,664 句的 NIST MT-06 當作是我們的測試資料。在這份資料中，共有散佈在 859 句的 933 種未知詞。細部分析顯示測試資料中未知詞個數和未知詞的長度關係和之前發展中資料是相當類似的：雙字（two-character）詞也佔了所有未知詞的一半。在這 933 種未知詞中，我們的未知詞模組爲 351 句中的 170 種雙字（two-character）未知詞產生翻譯候選，進而藉由底層 Moses 產生翻譯。我們系統中的門檻值 θ 決定了我們系統的應用程度，也就是，θ 被設計來藉由機率的大小稍加檢視組合式的翻譯是否適合該未知詞。模組所產生的翻譯候選將會利用 XML 標記加入 Moses 中。

表八：系統翻譯表現（句數 1,664）

翻譯系統	BLEU	BP	翻出字詞個數
Moses	21.46	0.928	41052
CST	21.56	0.939	41707
Fixed	21.34	0.941	41805

表九：系統翻譯表現（句數 351）

翻譯系統	BLEU	BP	翻出字詞個數
Moses	17.41	0.912	10833
CST	**17.83**	**0.951**	11583

表八整理出各系統翻譯的表現。雖然底層 Moses 和有加上我們模組的 Moses（命名爲系統 CST，因爲 Moses with combined sublexical translations）在 BLEU 的分數上並沒有很大的差異，但是 CST 在精簡懲罰（brevity penalty，也就是 BP）上則有明顯的上升，由此可知，CST 系統所產生出來的翻譯句長和參考翻譯（reference translation）的長度較爲接近。爲了檢驗 CST 系統所多產生的翻譯字詞的確如 BLEU 分數所顯示的一樣一比底層 Moses 的準確度更好或是至少一樣，我們多比較了 Fixed（表八最後一列）這個系統。爲了說明我們系統得到較高的 BLEU 分數除了因爲精簡懲罰較大（越大越好）以外，我們所多翻譯出來的字仍維持高正確性，我們將底層 Moses 未翻譯出來的雙字未知詞都以固定（Fixed）非中文字元帶入，並觀察其翻譯表現。誠如表八，分數下降的 Fixed 系統代表著即使翻譯長度近似參考翻譯的長度，沒有翻譯的準度，BLEU 的分數是不會上升的。也反映出 CST 系統爲雙字未知詞產生出不錯的翻譯。

我們更進一步來檢驗我們系統爲 351 句測試句產生未知詞翻譯的 BLEU 表現（請見表九）。CST 系統爲片語式翻譯系統 Moses 在 BLEU 分數上帶來的提升是數學統計上顯著的（statistically significant）。我們使用 Koehn 在 2004 年提到的 bootstrap resampling 方法來作顯著測試（significance test）。從表九翻譯分數較低（相較表八）顯示：這些句子較難翻譯，且很有可能是句子中未知詞的關係。另外，CST 系統在這些句子中的 BP 進步更大，比例（relatively）成長 4.4%之多。

總結的說，實驗數據顯示我們的未知詞組合模組可以翻譯部分的未知詞，且不會降低現存翻譯系統的表現。對於那些我們系統有產生翻譯候選的句子，翻譯表現是大幅的提升。

五、總結與未來展望

在本研究中，我們針對片語式機器翻譯系統的未知詞和落單字提出解決方案。在我們細部的分析中發現組成字（combinational form）未知詞的比例不亞於文獻較爲重視的可重述之未知詞比例。另外，片語式機器翻譯系統平均而言會忽略掉 5%以上的落單字翻譯品質，而落單字又是片語式翻譯系統沒有辦法（翻譯）解歧的對象（得藉助受字數限制的語言模組來幫忙）。這次的實驗我們專注在未知詞組合字翻譯模組的貢獻。此模組包含利用萬用字元查詢取的組成字翻譯、限制且過濾較不可能的組成字翻譯、組合組成字翻譯、並藉由單雙語的資訊來排序組合出來的翻譯。我們實驗結果是相當正面的：架構在知名的片語式機器翻譯系統 Moses 之上，未知詞組成字翻譯模組產生的翻譯候選清單很有可能就包含了未知詞的正確翻譯、組合式的翻譯未知詞可以有效地降低精簡懲罰（brevity penalty）、大大提升包含有未知詞句子翻譯的品質。我們的實驗結果也暗示所謂的中文未知詞在字層級（character-level）上可能是已知的（in-vocabulary）。

未來我們也希望我們可以將組合字模組拓展到可以翻譯三字（three-character）詞或是以上。例如：「國科會」或是「電視台」。然而，我們將會需要字層級的斷詞法。例如：「國科會」應該被切成「國」（national）、「科」（science）、和「會」（council），而「電視台」被切成「電視」（television）和「台」（station）。模組中的組成字的解歧也需要加強。例如：組成字「班」可能代表「航班」（flight）、「班車」（train）、「班級」（class）、和「值班」（shift）等。除了未知詞本身，我們也許可以利用其上下文來幫忙組成字解歧。另外，我們也將結合未知詞重述模組（例如：（Mirkin 等人, 2009）和（Marton 等人, 2009））以增加翻譯的涵蓋率。最後，雖然我們針對落單字所佔的比例還有各詞性 Moses 系統翻譯的準確度作了分析並提出屬於動詞和名詞的落單字是片語式機器翻譯的弱項，我們並沒有實際實驗我們的落單字翻譯模組的效用。未來我們將專注在發展實驗此翻譯模組上，並考慮合併此論文中提出來的組合字模組、重述模組、和落單字模組。

致謝辭

本研究依經濟部補助財團法人資訊工業策進會「100 年度數位匯流服務開放平台技術研發計畫」辦理。

參考文獻

Steven Bird, Ewan Klein, and Edward Loper. 2008. Natural language processing in Python. Available online at http://nltk.org/book.html.

Chu-Ren Huang, Ru-Yng Chang, and Shiang-Bin Lee. 2004. Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO. In *Proceedings of*

the Fourth International Conference on Language Resources and Evaluation (LREC), pages 1553-1556.

Philipp Koehn and Kevin Knight. 2003. Empirical Methods for Compound Splitting. In *Proceedings of the 10th conference on European chapter of the Association for Computational Linguistics*, pages 187-193.

Philipp Koehn. 2004. Statistical Significance Test for Machine Translation Evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388-395.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne and David Talbot. 2005. Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation. In *International workshop on Spoken Language Translation*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 177–180.

Wei-Yun Ma and Keh-Jiann Chen. 2003. Introduction to CKIP Chinese word segmentation system for the first international Chinese word segmentation bakeoff. In *Proceedings of the second SIGHAN workshop on Chinese language processing*, pages 168-171.

George A. Miller. 1995. Wordnet: A Lexical Database for English. *Communications of the ACM*, vol. 38, no. 11, pages 39-41.

Shachar Mirkin, Lucia Specia, Nicola Cancedda, Ido Dagan, Marc Dymetman, and Idan Szpektor. 2009. Source-language Entailment Modeling for Translating Unknown Terms. In *Proceedings of the 47th Annual Meeting of ACL and the 4th IJCNLP of the AFNLP*, pages 791–799.

Franz Josef Och and Hermann Ney. 2003. A systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, vol. 29, no. 1, pages 19-51.

Kishore Papineni, Salim Roukos, ToddWard, andWei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311-318.

Frank A. Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, vol. 19 (1), pages 143-177.

Andreas Stolcke. 2002. SRILM – An Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, vol. 2, pages 901–904.

Mei Yang and Katrin Kirchhoff. 2006. Phrase-Based Backoff Models for Machine Translation of Highly Inflected Languages. In *Proceedings of the European Chapter of the Association for Computational Linguistics*, pages 41-48.