: ---.

इंटरनेशनल साइंटफिकि जर्नल ऑफ़ इंजीनयिरगि एंड मैनेजमेंट () : 2583-
वॉल्यूम: 04 अंक: 05 |मई -
: 10.55041/03470
एक अंतरराष्ट्रीय वद्विवतापूर्ण ||बहु -वषियक ||ओपन एक्सेस ||सभी प्रमुख डेटाबेस और मेटाडेटा में अनुक्रमण

एआई-आधारति पीडीएफ अनुवादक
आर। राजेश*1, पी। अशोक*2, एम। साईं कृष्णा*
* इंजीनयिरगि कॉलेज, भारत के ( ) वभिाग के 1 सहायक प्रोफेसर।

* इंजीनयिरगि कॉलेज, भारत के ( ) वभिाग के 2,3 छात्र।

अमूर्त
स्वचालति दस्तावेज़ अनुवाद भाषा की बाधाओं पर काबू पाने और सीमलेस को सुवधिाजनक बनाने में महत्वपूर्ण भूमकिा नभिाता है
वैश्वकि उद्योगों में संचार।यह परयिोजना प्राकृतकि भाषा प्रसंस्करण (एनएलपी) की शक्तिका उपयोग करती है
और ऑप्टकिल चरत्रि मान्यता () पीडीएफ दस्तावेजों से पाठ को नकिालने, अनुवाद करने और पुनर्नर्माण करने के लएि
उनके मूल लेआउट और स्वरूपण को संरक्षति करना।ट्रांसफार्मर-आधारति मॉडल जैसे जीपीटी और
का उपयोग करके
एपीआई का अनुवाद करें, मजबूत पाठ नष्कर्षण उपकरणों के साथ, सस्टिम सटीक और कुशल बहुभाषी प्रदान करता है
अनुवाद।कार्यप्रणाली में , , -, और सहति पायथन पुस्तकालयों
को शामलि कयिा गया है
पाठ मान्यता, अनुवाद और सुधार प्रक्रयिाओं का प्रबंधन करने के लएि।यह एआई-संचालति
समाधान उद्देश्य है
पहुंच बढ़ाने के लएि, वैश्वकि सहयोग को बढ़ावा दें, और बहुभाषी दस्तावेज़ वर्कफ्लोज़ को सुव्यवस्थति करें
ववधि क्षेत्र।
कीवर्ड: पीडीएफ अनुवाद, प्राकृतकि भाषा प्रसंस्करण (एनएलपी), ऑप्टकिल चरत्रि मान्यता (ओसीआर), ट्रांसफार्मर मॉडल, बहुभाषी दस्तावेज़ प्रसंस्करण।

मैं।
परचिय
1.1. और प्रेरणा:
एक तेजी से वैश्वकि और परस्पर जुड़ी दुनयिा में, संगठनों को संवाद और वनिमिय करने की आवश्यकता होती है
भाषाई सीमाओं के पार जानकारी।पोर्टेबल डॉक्यूमेंट फॉर्मेट (पीडीएफ) में दस्तावेज़ व्यापक रूप से उपयोग कएि जाते हैं
स्वास्थ्य सेवा, वत्ति, कानून, शक्षिावदि और सरकार जैसे उद्योग उनकी पोर्टेबलिटी, सुसंगत स्वरूपण, के कारण,
और बहु-प्लेटफॉर्म संगतता।हालांकि, वभिन्नि भाषाओं में पीडीएफ दस्तावेजों का अनुवाद करना कई प्रस्तुत करता है
चुनौतयिां - वशिष रूप से जब इन दस्तावेजों में जटलि लेआउट, स्कैन की गई छवयिां और बहुभाषी सामग्री होती है।

पारंपरिक अनुवाद विधियों अक्सर मैनुअल, समय लेने वाली, महंगी और त्रुटियों के लिए प्रवण होती हैं, जिससे वे बनते हैं
बड़े पैमाने पर दस्तावेज़ प्रसंस्करण के लिए अक्षम।

1.2 परिचय
आज की तेजी से वैश्वीकृत दुनिया में, विभिन्न में जानकारी संवाद करने और साझा करने की क्षमता
भाषाएँ महत्वपूर्ण हैं।संगठन अक्सर विविध हितधारकों, भागीदारों और ग्राहकों के साथ काम करते हैं जो विभिन्न बोलते हैं
भाषाएं, कुशल, सटीक और स्केलेबल अनुवाद समाधानों के लिए बढ़ती मांग पैदा करती हैं।परंपरागत
दस्तावेज़ अनुवाद के तरीके समय लेने वाले, महंगे और मानवीय त्रुटि के लिए प्रवण हैं-विशेष रूप से व्यवहार करते समय
पीडीएफ जैसे जटिल प्रारूपों में सामग्री के बड़े संस्करणों के साथ।
पीडीएफ दस्तावेजों का उपयोग उनके निश्चित स्वरूपण और क्रॉस-प्लेटफॉर्म संगतता के कारण उद्योगों में व्यापक रूप से किया जाता है।
हालांकि, उनकी कठोर संरचना स्वचालित अनुवाद के लिए चुनौतियों का सामना करती है, खासकर जब वे स्कैन किए गए होते हैं
छवियां, टेबल, या बहुभाषी सामग्री।यह बुद्धिमान प्रौद्योगिकियों को अपनाने के लिए आवश्यक बनाता है जो सटीक रूप से कर सकते हैं
ऐसे प्रारूपों से पाठ्य डेटा निकालें, व्याख्या और पुन: इकट्ठा करें।
यह परियोजना स्वचालित पीडीएफ अनुवाद के लिए एआई-संचालित प्रणाली विकसित करके इन चुनौतियों को संबोधित करती है।द्वारा
लीवरेजिंग नेचुरल लैंग्वेज प्रोसेसिंग (एनएलपी), ऑप्टिकल कैरेक्टर रिकग्निशन (ओसीआर), और एडवांस्ड ट्रांसफार्मर-
और अनुवाद जैसे आधारित मॉडल, सिस्टम देशी और स्कैन किए गए पाठ दोनों का अनुवाद
कर सकता है जबकि

इंटरनेशनल साइंटिफिक जर्नल ऑफ़ इंजीनियरिंग एंड मैनेजमेंट () : 2583-
वॉल्यूम: 04 अंक: 05 |मई -
: 10.55041/03470
एक अंतरराष्ट्रीय विद्वतापूर्ण ||बहु -विषयक ||ओपन एक्सेस ||सभी प्रमुख डेटाबेस और मेटाडेटा में अनुक्रमण

मूल लेआउट को बनाए रखना।, , -, और जैसे पायथन
लाइब्रेरी
एपीआई इस समाधान की बैकबोन बनाता है, जो निष्कर्षण, अनुवाद और पुनर्निर्माण के सहज एकीकरण को सक्षम करता है
वर्कफ़्लोज़।
इस परियोजना का अंतिम लक्ष्य भाषा की बाधाओं को तोड़ना, पहुंच में सुधार करना और वैश्विक बढ़ाना है
बुद्धिमान बहुभाषी दस्तावेज़ प्रसंस्करण के माध्यम से सहयोग।
।साहित्यिक निरीक्षण
स्वचालित पीडीएफ अनुवाद प्रणालियों का विकास ऑप्टिकल चरित्र मान्यता में प्रगति को एकीकृत करता है
(), प्राकृतिक भाषा प्रसंस्करण (), और ट्रांसफार्मर-आधारित मॉडल।यह सर्वेक्षण प्रमुख
अनुसंधान पर प्रकाश डालता है
योगदान जो ऐसी प्रणालियों के डिजाइन और कार्यान्वयन को सूचित करते हैं।
2.1।पाठ निष्कर्षण के लिए इंजन
स्मिथ ने टेसरेक्ट को पेश किया, जो एक ओपन-सोर्स ओसीआर इंजन है जो विभिन्न भाषाओं में पाठ को पहचानने

में सक्षम है और
स्क्रिप्ट। की अनुकूलनशीलता और सटीकता इसे स्कैन किए गए दस्तावेजों से पाठ निकालने के
लिए एक मूल्यवान उपकरण बनाती है, ए
स्वचालित पीडीएफ अनुवाद वर्कफ़्लोज़ में महत्वपूर्ण कदम।

2.2।संयुक्त रूप से संरेखित करने और अनुवाद करने के लिए संयुक्त रूप से सीखने से तंत्रिका मशीन अनुवाद
बहडानाऊ एट अल।एक प्रारंभिक मॉडल प्रस्तुत किया जो संयुक्त रूप से संरेखित करना और अनुवाद करना
सीखता है, सीमाओं को संबोधित करता है
पारंपरिक अनुक्रम-से-अनुक्रम मॉडल।उनका दृष्टिकोण एक ध्यान तंत्र को एकीकृत करता है जो मॉडल को
सक्षम करता है
अनुवाद के दौरान स्रोत वाक्य के प्रासंगिक भागों पर ध्यान केंद्रित करने के लिए, लंबे अनुक्रमों पर प्रदर्शन में
सुधार।

2.3।ध्यान-आधारित तंत्रिका मशीन अनुवाद के लिए प्रभावी दृष्टिकोण
लुओंग एट अल।वैश्विक और स्थानीय सहित तंत्रिका मशीन अनुवाद में विभिन्न ध्यान तंत्रों का पता लगाया
ध्यान दें मॉडल।उनके निष्कर्षों से संकेत मिलता है कि ध्यान-आधारित मॉडल पारंपरिक रूप से बेहतर प्रदर्शन
करते हैं
अनुवाद प्रणाली, विशेष रूप से लंबे वाक्यों और जटिल संरचनाओं को संभालने में।

2.4।ध्यान आप सभी की जरूरत है (ट्रांसफार्मर वास्तुकला)
वासवानी एट अल। द्वारा फाउंडेशनल वर्क, शीर्षक "अटेंशन इज़ ऑल यू नीड," ने ट्रांसफार्मर का परिचय दिया
आर्किटेक्चर, क्रांति मशीन अनुवाद और एनएलपी कार्य।स्व-क्षरण तंत्र पर मॉडल की निर्भरता
समानांतर प्रसंस्करण के लिए अनुमति देता है और लंबी दूरी की निर्भरता के हैंडलिंग में सुधार करता है, नए
बेंचमार्क सेट करता है
अनुवाद की गुणवत्ता।

2.5। की तंत्रिका मशीन अनुवाद प्रणाली
वू एट अल।एक तंत्रिका मशीन अनुवाद प्रणाली के लिए के संक्रमण को विस्तृत करें जो गहरे
नेटवर्क का उपयोग करता है
ध्यान तंत्र।इस प्रणाली ने अनुवाद सटीकता और प्रवाह में पर्याप्त सुधार प्राप्त किया,
बड़े पैमाने पर अनुप्रयोगों में तंत्रिका दृष्टिकोण की व्यावहारिक व्यवहार्यता का प्रदर्शन।

2.6।बर्ट: भाषा समझ के लिए गहरी द्विदिशि ट्रांसफार्मर का पूर्व-प्रशिक्षण
डेवलिन एट अल।बर्ट का परिचय, एक पूर्व-प्रशिक्षित गहरी द्विदिशि ट्रांसफार्मर मॉडल जिसने अतृप्त है-
विभिन्न एनएलपी कार्यों में कला परिणाम।बर्ट की वास्तुकला दोनों दिशाओं में संदर्भ की गहरी समझ के लिए
अनुमति देती है,
प्रश्न उत्तर और भाषा के अनुमान जैसे कार्यों के लिए इसे अत्यधिक प्रभावी बनाना।

2.7।तंत्रिका अनुवाद मॉडल में व्याख्यात्मक ध्यान तंत्र
ज़ेनकेल एट अल।ट्रांसफार्मर आर्किटेक्चर के लिए एक एक्सटेंशन पेश किया जो शामिल करके शब्द संरेखण को
बढ़ाता है
व्याख्यात्मक ध्यान तंत्र।उनका मॉडल एनकोडर जानकारी पर ध्यान केंद्रित करता है, अधिक सुविधा प्रदान
करता है
प्रशिक्षण के दौरान शब्द-संरेखण डेटा की आवश्यकता के बिना सटीक संरेखण।यह दृष्टिकोण काफी है
ट्रांसफॉर्मर ध्यान सक्रियिणों की भोली व्याख्याओं को कम करता है और संरेखण गुणवत्ता को प्राप्त करता है
++ जैसे पारंपरिक उपकरणों के लिए।

2.8।ट्रांसफार्मर के लिए व्याख्या योग्य शब्द संरेखण
डौ और न्युबगि ने व्याख्याता को बढ़ाने के लिए हेड प्रूनगि और पर्यवेक्षण संरेखण तकनीकों की शुरुआत की

इंटरनेशनल साइंटफिकि जर्नल ऑफ़ इंजीनयिरगि एंड मैनेजमेंट () : 2583-
वॉल्यूम: 04 अंक: 05 |मई -
: 10.55041/03470
एक अंतरराष्ट्रीय वद्विवतापूर्ण ||बहु -वषियक ||ओपन एक्सेस ||सभी प्रमुख डेटाबेस और मेटाडेटा में अनुक्रमण

ट्रांसफार्मर में ध्यान तंत्र।उनका दृष्टकिोण कुंजी का चयन करके अधकि सटीक शब्द संरेखण देता है
ध्यान दें, हालांकि इसके लिए अतरिक्ति पर्यवेक्षण और संरेखण डेटा की आवश्यकता होती है।

2.9।ढाल-आधारति स्व-संयोग मानचत्रि वश्लिेषण
बार्कन एट अल।प्रस्तावति ग्रेड-एसएएम, ट्रांसफार्मर भवष्यिवाणयिों को स्वयं के माध्यम से समझाने के लिए एक ढाल-आधारति वधि
ध्यान दें।ग्रेड-एसएएम इनपुट तत्वों की पहचान करता है जो मॉडल नर्णियों को सबसे महत्वपूर्ण रूप से प्रभावति करते हैं, प्रदान करते हैं
ट्रांसफार्मर-आधारति भाषा मॉडल के आंतरकि कामकाज में अंतर्दृष्टि।मूल्यांकन से पता चलता है कि ग्रेड-सैम मौजूदा व्याख्या तकनीकों पर पर्याप्त सुधार प्रदान करता है।

2.10।बड़े भाषा मॉडल में अनुवाद तंत्र की खोज
झांग एट अल।बड़े भाषा मॉडल (एलएलएम) के भीतर अनुवाद तंत्र को समझने के लिए एक अध्ययन किया।
पथ पैचगि तकनीकों का उपयोग करते हुए, उन्होंने पाया कि ध्यान देने वाले सरि का एक वरिल सबसेट (5%से कम) मुख्य रूप से अनुवाद कार्यों की सुवधिा प्रदान करता है।केवल इन वशिेष सरि को ठीक करने के द्वारा, उन्होंने अनुवाद प्राप्त किया
पूर्ण-पैरामीटर ट्यूनगि की तुलना में सुधार, लक्षति मॉडल समायोजन की दक्षता को उजागर करते हुए।
2.11।तुलना तालकिा: ध्यान-आधारति मॉडल पर साहतिय समीक्षा
नहीं।
कागज का शीर्षक /
केंद्र
लेखक
वर्ष
कार्यप्रणाली प्रमुख नष्किर्ष
सीमाएँ
/
टप्पिणी

टेसरैक्ट
ओसीआर इंजन
लोहार

एलएसटीएम आधारति
ओसीआर
साथ
भाषा

सहायता
शुद्ध
बहुभाषी
मूलपाठ
मान्यता
छवियों से
संघर्ष
साथ
अत्यधिक शोर या
विषम
दस्तावेज़

तंत्रिका
मशीन
द्वारा अनुवाद करना
संयुक्त रूप से
सीखना
को
संरेखित
और
अनुवाद
बहदानौ
एट अल।

एनकोडर-
के साथ डिकोडर
ध्यान
तंत्र
उन्नत
का अनुवाद
लंबा
दृश्यों
कम्प्यूटेशनल के रूप में
गैर से धीमा
ध्यान
2 मॉडल

असरदार
के लिए दृष्टिकोण
ध्यान-
आधारित तंत्रिका
मशीन
अनुवाद

एट
अल।

वैश्वकि
बनाम
स्थानीय
ध्यान
में
2
मॉडल
स्थानीय
ध्यान
तेजी से मॉडल
और
अधकि
शुद्ध
में
नश्चिति
संदर्भों
आरएनएन तक सीमति
चौखटे

:
'
तंत्रकि
मशीन
अनुवाद
प्रणाली
वू एट अल।

गहरी
ध्यान के साथ
के लिए
अंत हैं-
अंत एमटी
प्रमुख
सुधार
प्रवाह में और
अनुवाद
पैमाने पर गुणवत्ता
उच्च
वलिंब
बनिा
अनुकूलन

ध्यान
है
आप सभी की जरूरत
वासवानी एट
अल।

ट्रांसफार्मर
वास्तुकला
का उपयोग करते हुए
खुद-
ध्यान
पेश किया
ट्रांसफार्मर;
बेहतर प्रदर्शन किया

में
अनुवाद
कार्य
आवश्यक है
बड़ा
डेटासेट
और
गणना
संसाधन

बर्ट:
द्विदिशि
ट्रान्सफ़ॉर्मर
डेवलनि
एट
अल।

पूर्व प्रशिक्षण
नकाबपोश के साथ
भाषा
मॉडलिंग
मज़बूत
प्रासंगिक
समझ;
नहीं
अनुकूलित
के लिए
अनुवाद
कार्य
विशेष रूप से

इंटरनेशनल साइंटिफिक जर्नल ऑफ़ इंजीनियरिंग एंड मैनेजमेंट () : 2583-
वॉल्यूम: 04 अंक: 05 |मई -
: 10.55041/03470
एक अंतरराष्ट्रीय विद्वतापूर्ण ||बहु -विषयक ||ओपन एक्सेस ||सभी प्रमुख डेटाबेस और मेटाडेटा में
अनुक्रमण

भाषा के लिए
समझ
बेहतर एनएलपी
मानक

इन्तेर्प्रेतिंग
ध्यान
में
ट्रांसफार्मर
मॉडल
ज़ेनकेल
एट
अल।

विशेष
ध्यान
को
बढ़ाना
संरेखण
प्राप्त
++-स्तर
संरेखण
बिना
बाहरी
संरेखण आंकड़ा
सामान्यीकृत नहीं
व्यापक एनएलपी को
कार्य

व्याख्या
शब्द
के लिए संरेखण
ट्रान्सफ़ॉर्मर
डोऊ
और
न्यूबिंग

हेड प्रूनिंग
और
देखरेख
संरेखण
सिग्नल
का उत्पादन
व्याख्या
और
शुद्ध

शब्द
संरेखण
आवश्यक है
अतिरिक्त
पर्यवेक्षण

ग्रेड-एसएएम:
तस्वीर
स्पष्टीकरण
के लिए
ट्रांसफार्मर
फैसले
बर्कन एट
अल।

ग्रेडिएंट-
भारित
आत्मनिर्णय
एमएपीएस
ऑफर
उन्नत
विवेचनीयता
के लिए
नमूना
फैसले
ध्यान केंद्रित
प्रमुख रूप से
पर
स्पष्टता

कैसे
करना
लल्म्स
अनुवाद करना?
इन्तेर्प्रेतिंग
साथ
पथ
पैच
झांग एट अल।
पाथ पैचिंग
ध्यान के लिए
प्रवाह अनुरेखण
केवल ~ 5%
ध्यान
प्रमुख कुंजी हैं
अनुवाद में
कार्य

अनुवाद में
ध्यान केंद्रित,
का अभाव
सामान्य
आंदोलन
अंतर्दृष्टि

।अनुसंधान अंतराल
ध्यान-आधारित मॉडल और बहुभाषी दस्तावेज़ प्रसंस्करण में महत्वपूर्ण प्रगति के बावजूद, कई महत्वपूर्ण अनुसंधान अंतराल बने हुए हैं।सबसे पहले, जबकि ट्रांसफार्मर-आधारित मॉडल जैसे कि बर्ट और जीपीटी उच्च गुणवत्ता की पेशकश करते हैं
अनुवाद, वे अक्सर मूल लेआउट को संरक्षित करने और जटिल पीडीएफ दस्तावेजों के स्वरूपण की क्षमता में कमी करते हैं।
यह सीमा कानूनी, शैक्षणिक या वित्तीय दस्तावेजों जैसे संदर्भों में महत्वपूर्ण है, जहां संरचना बताती है अर्थ।दूसरा, अधिकांश अनुवाद प्रणाली क्लाउड-आधारित एपीआई पर बहुत अधिक भरोसा करती है, गोपनीयता की चिंताओं को पूरा करती है और सीमित करती है
ऑफ़लाइन या सुरक्षित वातावरण में प्रयोज्यता।तीसरा, हालांकि विभिन्न अध्ययनों ने मॉडल को बढ़ाया है ध्यान विज़ुअलाइज़ेशन और ग्रेडिएंट विश्लेषण का उपयोग करते हुए, अभी भी सीमित समझ है कि कैसे विशिष्ट ध्यान प्रमुख विभिन्न वाक्य रचना संरचनाओं के साथ भाषाओं में अनुवादों को प्रभावित करते हैं।इसके अतिरिक्त,
मौजूदा उपकरण जैसे कम-रिज़ॉल्यूशन या मल्टी-कॉलम पीडीएफ के साथ संघर्ष करते हैं,
जो डाउनस्ट्रीम को प्रभावित करता है
अनुवाद सटीकता।एक अन्य प्रमुख अंतर एंड-टू-एंड ओपन-सोर्स सिस्टम की अनुपस्थिति है जो ओसीआर, एनएलपी को एकीकृत करता है,
एक एकीकृत, मॉड्यूलर पाइपलाइन में अनुवाद, और लेआउट पुनर्निर्माण।अंत में, कई मॉडलों को प्रशिक्षित किया गया है
मुख्य रूप से उच्च-संसाधन वाली भाषाओं पर, दोनों के संदर्भ में कम-संसाधन वाली भाषा जोड़े को कम करना अनुवाद गुणवत्ता और मॉडल मूल्यांकन।इन अंतरालों को संबोधित करने से अधिक मजबूत, व्याख्या करने योग्य का मार्ग प्रशस्त हो सकता है,
और समावेशी बहुभाषी दस्तावेज़ अनुवाद प्रणाली।
।प्रस्तावित कार्यप्रणाली

प्रस्तावित प्रणाली
बहुभाषी पीडीएफ दस्तावेजों को निकालने, अनुवाद करने और पुनर्निर्माण के लिए एआई-संचालित पाइपलाइन विकसित करना है
उनके मूल स्वरूपण और संरचना को संरक्षित करते हुए।कार्यप्रणाली ऑप्टिकल चरित्र मान्यता को एकीकृत करती है
(), नेचुरल लैंग्वेज प्रोसेसिंग (), और ट्रांसफॉर्मर-आधारित ट्रांसलेशन मॉडल एक मॉड्यूलर फ्रेमवर्क में।
प्रक्रिया को पांच प्रमुख चरणों में विभाजित किया गया है:
3.1।पीडीएफ पाठ निष्कर्षण

 और का उपयोग करना
, सिस्टम पहले फ़ाइलों से मशीन-पठनीय और छवि-आधारित पाठ दोनों का पता लगाता है
और निकालता है।के लिए

इंटरनेशनल साइंटफिकि जर्नल ऑफ़ इंजीनयिरगि एंड मैनेजमेंट () : 2583-
वॉल्यूम: 04 अंक: 05 |मई -
: 10.55041/03470

स्कैन या छवि-आधारति दस्तावेज़, - को छवयिों को संपादन योग्य पाठ में परविर्तति करने के
लएि नयिोजति कयिा जाता है जबकि
पाठ पदों, स्तंभों और पैराग्राफ की पहचान करना।
3.2।प्रीप्रोसेसगि और भाषा का पता लगाना

नकिाले
मूलपाठ
है
साफ, तार्ककि ब्लॉकों में खंडति (जैसे, शीर्षकों, पैराग्राफ, टेबल), और भाषा का पता लगाने के माध्यम से पारति कयिा गया
स्रोत भाषा की पहचान करने के लएि एल्गोरदिम (जैसे,)।यह कदम सटीक भाषा-वशिष्टि
सुनश्चिति करता है
अनुवाद हैंडलगि।
3.3।मशीन अनुवाद

पूर्वप्रभरण कयिा हुआ
पाठ का अनुवाद ट्रांसफार्मर-आधारति मॉडल जैसे कि या अनुवाद का उपयोग
करके कयिा जाता है, जो इस पर नर्भिर करता है
परनियिोजन संदर्भ।सस्टिम लेआउट अखंडता को बनाए रखने के लएि टेक्स्ट ब्लॉक-बाय-ब्लॉक अनुवाद को संभालता है।वकिसति
मॉडल डोमेन-वशिष्टि भाषा उपयोग और बढ़ाया प्रवाह के लएि ठीक-ठीक हो सकते हैं।
3.4।लेआउट संरक्षण और पुनर्नर्माण

का उपयोग करते हुए

मेटाडेटा मूल दस्तावेज़ से (नर्देशांक, फ़ॉन्ट आकार, स्थतिि), अनुवादति पाठ को एक नए में फरि से इंजेक्ट कयिा गया है
पीडीएफ लेआउट। या जैसे पुस्तकालय मूल दस्तावेज़ की संरचना को दोहराने में मदद
करते हैं, जसिमें शामलि हैं
छवयिां, टेबल और स्वरूपण।
3.5।पोस्टप्रोसेसगि और मूल्यांकन

आउटपुट
पीडीएफ एक गुणवत्ता आश्वासन चरण से गुजरता है जसिमें समिंटकि स्थरिता जांच, लेआउट सत्यापन और शामलि हैं
या उल्का स्कोर का उपयोग करके भाषा की गुणवत्ता का मूल्यांकन।वैकल्पकि मानव-इन-द-लूप सत्यापन हो सकता है
उच्च-दांव अनुप्रयोगों के लएि एकीकृत।
वी। नष्किर्ष
यह

परियोजना,
हमने बहुभाषी पीडीएफ अनुवाद के लिए एक एआई-आधारित प्रणाली का प्रस्ताव दिया जो अत्याधुनिक प्रौद्योगिकियों को एकीकृत करता है
ऑप्टिकल चरित्र मान्यता (), प्राकृतिक भाषा प्रसंस्करण (), और ट्रांसफार्मर-आधारित
मॉडल।द्वारा
-, , और और जैसे शक्तिशाली अनुवाद मॉडल जैसे
टूल का लाभ उठाना
एपीआई का अनुवाद करें, सिस्टम कुशलता से अर्क, अनुवाद करता है, और बनाए रखते समय जटिल दस्तावेजों का पुनर्निर्माण करता है
मूल लेआउट और संरचना।
प्रस्तावित कार्यप्रणाली बहुभाषी दस्तावेज़ प्रसंस्करण में महत्वपूर्ण चुनौतियों को संबोधित करती है, जिसमें पाठ निष्कर्षण भी शामिल है
मशीन-पठनीय और छवि-आधारित पीडीएफ दोनों से, कई भाषाओं में सटीक अनुवाद, और
स्वरूपण, तालिकाओं और छवियों का संरक्षण।इसके अतिरिक्त, सिस्टम की मॉड्यूलर और स्केलेबल प्रकृति अनुमति देता है
अन्य दस्तावेज़ प्रसंस्करण वर्कफ़्लोज़ के साथ आसान एकीकरण के लिए और इसकी क्षमताओं को अलग करने की क्षमता को अलग -अलग करने की क्षमता
भाषा और डोमेन।
हालांकि, भविष्य में वृद्धि के लिए अभी भी कई क्षेत्र हैं, जैसे कि कम-रिज़ॉल्यूशन के लिए ओसीआर सटीकता में सुधार करना
दस्तावेज, कम-संसाधन भाषाओं के लिए अनुवाद की गुणवत्ता का अनुकूलन, और मॉडल व्याख्याता को बढ़ाने।
कुल मिलाकर, यह परियोजना दस्तावेज़ प्रसंस्करण में भाषा की बाधाओं को तोड़ने के लिए एक व्यापक समाधान प्रदान करती है,
बेहतर पहुंच और वैश्विक सहयोग को बढ़ावा देना।
।प्रतिक्रिया दें संदर्भ
[१]।वासवानी, ए।, शेज़ीर, एन।, परमार, एन।, उस्ज़कोरिट, जे।, जोन्स, एल।, गोमेज़, ए।, कैसर, ।, और पोलोसुखिन, आई। (2017)।
"ध्यान आप सभी की जरूरत है।"न्यूरिप्स 2017 की कार्यवाही।
[२]।बहडानौ, डी।, चो, के।, और बेंगियो, वाई। (2014)।"संयुक्त रूप से संरेखित करने के लिए और संयुक्त रूप से सीखने के द्वारा तंत्रिका मशीन अनुवाद
अनुवाद। " 2015 की कार्यवाही।
[३]।लुओंग, एम। टी।, फाम, एच।, और मैनिंग, सी। डी। (2015)।"ध्यान-आधारित तंत्रिका मशीन के लिए प्रभावी दृष्टिकोण
अनुवाद। " 2015 की कार्यवाही।

इंटरनेशनल साइंटिफिक जर्नल ऑफ़ इंजीनियरिंग एंड मैनेजमेंट () : 2583-
वॉल्यूम: 04 अंक: 05 |मई -
: 10.55041/03470
एक अंतरराष्ट्रीय विद्वतापूर्ण ||बहु -विषयक ||ओपन एक्सेस ||सभी प्रमुख डेटाबेस और मेटाडेटा में अनुक्रमण

[४]।ज़ेनकेल, एस।, गेहरिंग, जे।, और औली, एम। (2019)।"ट्रांसफार्मर मॉडल में ध्यान आकर्षित करना।"की कार्यवाही
एसीएल 2019।
[५]।बार्कन, एस।, डागन, आई।, और शवार्ट्ज, वी। (2022)।"ग्रेड-एसएएम: ट्रांसफार्मर निर्णयों के लिए दृश्य स्पष्टीकरण।"

2022 की कार्यवाही।
[६]।झांग, एक्स।, ज़ेंग, एक्स।, और ली, एस। (2024)।"एलएलएम कैसे अनुवाद करते हैं? पथ पैचिंग के साथ व्याख्या करना।"कार्यवाही
2024 का।
[[]।डौ, जेड।, और न्यूबिग, जी। (2021)।"ट्रांसफार्मर के लिए व्याख्या योग्य शब्द संरेखण।" 2021 की कार्यवाही।
[[]।स्मिथ, आर। (2007)।" इंजन।" 2007 की कार्यवाही।
[९]।वू, वाई।, शूस्टर, एम।, चेन, जेड।, ले।":
की न्यूरल मशीन अनुवाद प्रणाली। " 2016 की कार्यवाही।
[१०]।डेवलिन, जे।, चांग, एम। डब्ल्यू।, ली, के।, और टाउटनोवा, के। (2018)।"बर्ट: भाषा के लिए द्विदिशि ट्रांसफार्मर
समझ। " 2019 की कार्यवाही।